(51) International Patent Classification [6] :		(11) International Publication Number:	**WO 97/15690**
C12Q 1/68, 1/70, C12P 19/34, C12M 1/34, G01N 27/00, 21/64, C07H 21/02, 21/04, G06F 17/00	A1	(43) International Publication Date:	1 May 1997 (01.05.97)

(21) International Application Number: PCT/US96/17159

(22) International Filing Date: 24 October 1996 (24.10.96)

(30) Priority Data:
547,214   24 October 1995 (24.10.95)   US
663,823   14 June 1996 (14.06.96)   US

(71) Applicant: CURAGEN CORPORATION [US/US]; 322 East Main Street, Branford, CT 06405 (US).

(72) Inventors: ROTHBERG, Jonathan, M.; 45B Cocheco Avenue, Branford, CT 06405 (US). DEEM, Michael, W.; 11136 Ophir Drive, Los Angeles, CA 90024 (US). SIMPSON, John, W.; Apartment B4, 23 Woodland Road, Madison, CT 06443 (US).

(74) Agents: ANTLER, Adriane, M. et al.; Pennie & Edmonds, 1155 Avenue of the Americas, New York, NY 10036 (US).

(81) Designated States: AL, AM, AU, AZ, BA, BB, BG, BR, BY, CA, CN, CU, CZ, EE, FI, GE, HU, IL, IS, JP, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LV, MD, MG, MK, MN, MX, NO, NZ, PL, RO, RU, SG, SI, SK, TJ, TM, TR, TT, UA, UZ, VN, ARIPO patent (KE, LS, MW, SD, SZ, UG), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).

Published
With international search report.

(54) Title: METHOD AND APPARATUS FOR IDENTIFYING, CLASSIFYING, OR QUANTIFYING DNA SEQUENCES IN A SAMPLE WITHOUT SEQUENCING

(57) Abstract

This invention provides methods by which biologically derived DNA sequences in a mixed sample or in an arrayed single sequence clone can be determined and classified without sequencing. The methods make use of information on the presence of carefully chosen target subsequences, typically of length from 4 to 8 base pairs, and preferably the length between target subsequences in a sample DNA sequence together with DNA sequence databases containing lists of sequences likely to be present in the sample to determine a sample sequence. The preferred method uses restriction endonucleases to recognize target subsequences and cut the sample sequence. Then carefully chosen recognition moieties are ligated to the cut fragments, the fragments amplified, and the experimental observation made. Polymerase chain reaction (PCR) is the preferred method of amplification. Several alternative embodiments are described which are capable of increased discrimination and which use TypeIIS restriction endonucleases, various capture moieties, or samples of specially synthesized cDNA. Another embodiment of the invention uses information on the presence or absence of carefully chosen target subsequences in a single sequence clone together with DNA sequence databases to determine the clone sequence. Computer implemented methods are provided to analyze the experimental results and to determine the sample sequences in question and to carefully choose target subsequences in order that experiments yield a maximum amount of information.

METHOD AND APPARATUS FOR IDENTIFYING, CLASSIFYING, OR QUANTIFYING DNA SEQUENCES IN A SAMPLE WITHOUT SEQUENCING

The application is a continuation-in-part of
5 copending U.S. Patent Application serial number 08/547,214,
filed on October 24, 1995, which is hereby incorporated by
reference in its entirety.

This invention was made with United States
Government support under award number 70NANB5H1036 awarded by
10 the National Institute of Standards and Technology. The
United States Government has certain rights in the invention.

1. FIELD OF THE INVENTION

The field of this invention is DNA sequence
15 classification, identification or determination, and
quantification; more particularly it is the quantitative
classification, comparison of expression, or identification
of preferably all DNA sequences or genes in a sample without
performing any sequencing.
20

2. BACKGROUND

Over the past ten years, as biological and genomic
research have revolutionized our understanding of the
molecular basis of life, it has become increasingly clear
25 that the temporal and spatial expression of genes is
responsible for all life's processes, processes occurring in
both health and in disease. Science has progressed from an
understanding of how single genetic defects cause the
traditionally recognized hereditary disorders, such as the
30 thalassemias, to a realization of the importance of the
interaction of multiple genetic defects along with
environmental factors in the etiology of the majority of more
complex disorders, such as cancer. In the case of cancer,
current scientific evidence demonstrates the key causative
35 roles of alt red expression of and multiple defects in
several pivotal gen s. Other complex diseases have similar

etiology. Thus the more complete and reliable a correlation
that can be established between gene xpression and h alth or
disease states, the better diseases can be recognized,
diagnosed and treated.

5   This important correlation is established by the
quantitative determination and classification of DNA
expression in tissue samples, and such a method which is
rapid and economical would be of considerable value. Genomic
DNA ("gDNA") sequences are those naturally occurring DNA
10 sequences constituting the genome of a cell. The state of
gene, or gDNA, expression at any time is represented by the
composition of total cellular messenger RNA ("mRNA"), which
is synthesized by the regulated transcription of gDNA.
Complementary DNA ("cDNA") sequences are synthesized by
15 reverse transcription from mRNA. cDNA from total cellular
mRNA also represents, albeit approximately, gDNA expression
in a cell at a given time. Consequently, rapid and
economical detection of all the DNA sequences in particular
cDNA or gDNA samples is desired, particularly so if such
20 detection was rapid, precise, and quantitative.

Heretofore, gene specific DNA analysis techniques
have not been directed to the determination or classification
of substantially all genes in a DNA sample representing total
cellular mRNA and have required some degree of sequencing.
25 Generally, existing cDNA, and also gDNA, analysis techniques
have been directed to the determination and analysis of one
or two known or unknown genetic sequences at one time. These
techniques have used probes synthesized to specifically
recognize by hybridization only one particular DNA sequence
30 or gene. (See, e.g., Watson et al., 1992, Recombinant DNA,
chap 7, W. H. Freeman, New York.) Further, adaptation of
these methods to the problem of recognizing all sequences in
a sample would be cumbersome and uneconomical.

One existing method for finding and sequencing
35 unknown genes starts from an arrayed cDNA library. From a
particular tissue or specim n, mRNA is isolated and cloned
into an appropriate vector, which is then plat d in a manner

so that the progeny of individual vectors b aring the clone
of one cDNA sequence can be separately identified. A replica
of such a plate is then probed, often with a labeled DNA
oligomer selected to hybridize with the cDNA representing the
5 gene of interest. Thereby, those colonies bearing the cDNA
of interest are found and isolated, the cDNA harvested and
subject to sequencing. Sequencing can then be done by the
Sanger dideoxy chain termination method (Sanger et al., 1977,
"DNA sequencing with chain terminating inhibitors", Proc.
10 Natl. Acad. Sci. USA 74(12):5463-5467) applied to inserts so
isolated.

        The DNA oligomer probes for the unknown gene used
for colony selection are synthesized to hybridize,
preferably, only with the cDNA for the gene of interest. One
15 manner of achieving this specificity is to start with the
protein product of the gene of interest. If a partial
sequence of 5 to 10-mer peptide fragment from an active
region of this protein can be determined, corresponding 15 to
30-mer degenerate oligonucleotides can be synthesized which
20 code for this peptide. This collection of degenerate
oligonucleotides will typically be sufficient to uniquely
identify the corresponding gene. Similarly, any information
leading to 15 to 30 long nucleotide subsequences can be used
to create a single gene probe.

25   Another existing method, which searches for a known
gene in a cDNA or gDNA prepared from a tissue sample, also
uses single gene or single sequence probes which are
complementary to unique subsequences of the already known
gene sequences. For example, the expression of a particular
30 oncogene in sample can be determined by probing tissue
derived cDNA with a probe derived from a subsequence of the
oncogene's expressed sequence tag. Similarly the presence of
a rare or difficult to culture pathogen, such as the TB
bacillus or the HIV, can be determined by probing gDNA with a
35 hybridization probe specific to a gene of the pathogen. Th
heterozygous presence of a mutant allele in a phenotypically
normal individual, or its hom zygous presenc in a fetus, can

be determined by probing with an allele sp cific probe
complementary only to the mutant allele (See, e.g., Guo et
al., 1994, Nucleic Acid Research, 22:5456-65).

5 All existing methods using single gene probes, of
which the preceding examples are typical, if applied to
determine all genes expressed in a given tissue sample, would
require many thousands to tens of thousands of individual
probes. It is estimated a single human cell typically
expresses approximately to 15,000 to 15,000 genes
10 simultaneously and that the most complex tissue, e.g., the
brain, can express up to half the human genome (Liang et al.,
1992, "Differential Display of Eukaryotic Messenger RNA by
Means of the Polymerase Chain Reaction, Science, 257:967-
971). Such an application requiring such a number of probes
15 is clearly too cumbersome to be economic or, even, practical.

Another class of existing methods, known as
sequencing by hybridization ("SBH"), in contrast, use
combinatorial probes which are not gene specific (Drmanac et
al., 1993, Science 260:1649-52; U.S. Patent No. 5,202,231,
20 Apr 13, 1993, to Drmanac et al). An exemplary implementation
of SBH to determine an unknown gene requires that a single
cDNA clone be probed with all DNA oligomers of a given
length, say, for example, all 6-mers. Such a set of all
oligomers of a given length synthesized without any selection
25 is called a combinatorial probe library. From knowledge of
all hybridization results for a combinatorial library, say
all the 4096 6-mer probe results, a partial DNA sequence for
the cDNA clone can be reconstructed by algorithmic
manipulations. Complete sequences are not determinable
30 because, at least, repeated subsequences cannot be fully
determined. SBH adapted to the classification of known genes
is called oligomer sequence signatures ("OSS") (Lennon et
al., 1991, Trends In Genetics 7(10):314-317). This technique
classifies a single clone based on the pattern of probe hits
35 against an entire combinatorial library, or a significant
sub-library. It r quir s that the tissue sample library be

arrayed into clones, each clone comprising only one pure
sequence from the library. It cannot be applied to mixtures.
          These exemplary existing methods are all directed
to finding one sequence in an array of clones each expressing
5 a single sequence from a tissue sample. They are not
directed to rapid, economical, quantitative, and precise
characterization of all the DNA sequences in a mixture of
sequences, such as a particular total cellular cDNA or gDNA
sample. Their adaptation to such a task would be
10 prohibitive. Determination by sequencing the DNA of a clone,
much less an entire sample of thousands of sequences, is not
rapid or inexpensive enough for economical and useful
diagnostics. Existing probe-based techniques of gene
determination or classification, whether the genes are known
15 or unknown, require many thousands of probes, each specific
to one possible gene to be observed, or at least thousands or
even tens of thousands of probes in a combinatorial library.
Further, all of these methods require the sample be arrayed
into clones each expressing a single gene of the sample.
20   In contrast to the prior exemplary existing gene
determination and classification techniques, another existing
technique, known as differential display, attempts to
fingerprint a mixture of expressed genes, as is found in a
pooled cDNA library. This fingerprint, however, seeks merely
25 to establish whether two samples are the same or different.
No attempt is made to determine the quantitative, or even
qualitative, expression of particular, determined genes
(Liang et al., 1995, *Current Opinions in Immunology* 7:274-
280; Liang et al., 1992, *Science* 257:967-71; Welsh et al.,
30 1992, *Nucleic Acid Res.* 20:4965-70; McClelland et al., 1993,
Exs 67:103-15; Lisitsyn, 1993, *Science* 259:946-50).
Differential display uses the polymerase chain reaction
("PCR") to amplify DNA subsequences of various lengths, which
are defined by being between the hybridization sites of
35 arbitrarily select d primers. Ideally, the pattern of
lengths observed is characteristic of the tissue from which
the library was pr pared. Typically, one primer used in

diff rential display is oligo(dT) and the other is one r
more arbitrary oligonucleotides design d to hybridize within
a few hundred base pairs of the poly-dA tail of a cDNA in the
library. Thereby, on electrophoretic separation, the
5 amplified fragments of lengths up to a few hundred base pairs
should generate bands characteristic and distinctive of the
sample. Changes in tissue gene expression may be observed as
changes in one or more bands.

Although characteristic banding patterns develop,
10 no attempt is made to link these patterns to the expression
of particular genes. The second arbitrary primer cannot be
traced to a particular gene. First, the PCR process is less
than ideally specific. One to a few base pair ("bp")
mismatches ("bubbles") are permitted by the lower stringency
15 annealing step typically used and are tolerated well enough
so that a new chain can be initiated by the *Taq* polymerase,
often used in PCR reactions. Second, the location of a
single subsequence or its absence is insufficient infcrmation
to distinguish all expressed genes. Third, length
20 information from the arbitrary primer to the poly-dA tail is
generally not found to be characteristic of a sequence due to
variations in the processing of the 3' untranslated regions
of genes, the variation in the poly-adenylation process and
variability in priming to the repetitive sequence at a
25 precise point. Thus, even the bands that are produced often
are smeared by the non-specific background sequences present.
Also known PCR biases to high G+C content and short sequences
further limit the specificity of this method. Thus this
technique is generally limited to "fingerprinting" samples
30 for a similarity or dissimilarity determination and is
precluded from use in quantitative determination of the
differential expression of identifiable genes.

Existing methods for gene or DNA sequence
classification or determination are in need of improvement in
35 their ability to perform rapid and economical as well as
quantitative and specific determination of th components of
a cDNA mixture prepared from a tissue sample. Th preceding

background review identifies the deficiencies of several
exemplary existing methods.

3. SUMMARY OF THE INVENTION

5   It is an object of this invention to provide
methods for rapid, economical, quantitative, and precise
determination or classification of DNA sequences, in
particular genomic or complementary DNA sequences, in either
arrays of single sequence clones or mixtures of sequences
10 such as can be derived from tissue samples, without actually
sequencing the DNA. Thereby, the deficiencies in the
background arts just identified are solved. This object is
realized by generating a plurality of distinctive and
detectable signals from the DNA sequences in the sample being
15 analyzed. Preferably, all the signals taken together have
sufficient discrimination and resolution so that each
particular DNA sequence in a sample may be individually
classified by the particular signals it generates, and with
reference to a database of DNA sequences possible in the
20 sample, individually determined. The intensity of the
signals indicative of a particular DNA sequence depends
quantitatively on the amount of that DNA present.
Alternatively, the signals together can classify a
predominant fraction of the DNA sequences into a plurality of
25 sets of approximately no more than two to four individual
sequences.

           It is a further object that the numerous signals be
generated from measurements of the results of as few a number
of recognition reactions as possible, preferably no more than
30 approximately 5-400 reactions, and most preferably no more
than approximately 20-50 reactions. Rapid and economical
determinations would not be achieved if each DNA sequence in
a sample containing a complex mixture required a separate
reaction with a unique probe. Preferably, each recognition
35 reaction gen rates a large number of or a distinctive pattern
of distinguishable signals, which are quantitatively
proportional to the amount of the particular DNA sequences

present. Further, th signals are preferably detected and
measured with a minimum number of observations, which are
preferably capable of simultaneous performance.

The signals are preferably optical, generated by
5 fluorochrome labels and detected by automated optical
detection technologies. Using these methods, multiple
individually labeled moieties can be discriminated even
though they are in the same filter spot or gel band. This
permits multiplexing reactions and parallelizing signal
10 detection. Alternatively, the invention is easily adaptable
to other labeling systems, for example, silver staining of
gels. In particular, any single molecule detection system,
whether optical or by some other technology such as scanning
or tunneling microscopy, would be highly advantageous for use
15 according to this invention as it would greatly improve
quantitative characteristics.

According to this invention, signals are generated
by detecting the presence (hereinafter called "hits") or
absence of short DNA subsequences (hereinafter called
20 "target" subsequences) within a nucleic acid sequence of the
sample to be analyzed. The presence or absence of a
subsequence is detected by use of recognition means, or
probes, for the subsequence. The subsequences are recognized
by recognition means of several sorts, including but not
25 limited to restriction endonucleases ("REs"), DNA oligomers,
and PNA oligomers. REs recognize their specific subsequences
by cleavage thereof; DNA and PNA oligomers recognize their
specific subsequences by hybridization methods. The
preferred embodiment detects not only the presence of pairs
30 of hits in a sample sequence but also include a
representation of the length in base pairs between adjacent
hits. This length representation can be corrected to true
physical length in base pairs upon removing experimental
biases and errors of the length separation and detection
35 means. An alternative embodiment detects only the pattern of
hits in an array of clones, each containing a single sequ nce
("single s quence clon s").

- 8 -

The generated signals are th n analyzed together
with DNA sequence information stored in sequence databases in
computer implemented experimental analysis methods of this
invention to identify individual genes and their quantitative
5 presence in the sample.

The target subsequences are chosen by further
computer implemented experimental design methods of this
invention such that their presence or absence and their
relative distances when present yield a maximum amount of
10 information for classifying or determining the DNA sequences
to be analyzed. Thereby it is possible to have orders of
magnitude fewer probes than there are DNA sequences to be
analyzed, and it is further possible to have considerably
fewer probes than would be present in combinatorial libraries
15 of the same length as the probes used in this invention. For
each embodiment, target subsequences have a preferred
probability of occurrence in a sequence, typically between 5%
and 50%. In all embodiments, it is preferred that the
presence of one probe in a DNA sequence to be analyzed is
20 independent of the presence of any other probe.

Preferably, target subsequences are chosen based on
information in relevant DNA sequence databases that
characterize the sample. A minimum number of target
subsequences may be chosen to determine the expression of all
25 genes in a tissue sample ("tissue mode"). Alternatively, a
smaller number of target subsequences may be chosen to
quantitatively classify or determine only one or a few
sequences of genes of interest, for example oncogenes, tumor
suppressor genes, growth factors, cell cycle genes,
30 cytoskeletal genes, etc ("query mode").

A preferred embodiment of the invention, named
quantitative expression analysis ("QEA™"), produces signals
comprising target subsequence presence and a representation
of the length in base pairs along a gene between adjacent
35 target subsequences by m asuring the results of recognition
reactions on cDNA (or gDNA) mixtures. Of gr at importance,
this method does not require the cDNA be inserted into a

vector to create individual clones in a library. Creati n of thes libraries is time consuming, costly, and introduces bias into the process, as it requires the cDNA in the vector to be transformed into bacteria, the bacteria arrayed as
5 clonal colonies, and finally the growth of the individual transformed colonies.

Three exemplary experimental methods are described herein for performing QEA™: a preferred method utilizing a novel RE/ligase/amplification procedure; a PCR based method;
10 and a method utilizing a removal means, preferably biotin, for removal of unwanted DNA fragments. The preferred method generates precise, reproducible, noise free signatures for determining individual gene expression from DNA in mixtures or libraries and is uniquely adaptable to automation, since
15 it does not require intermediate extractions or buffer exchanges. A computer implemented gene calling step uses the hit and length information measured in conjunction with a database of DNA sequences to determine which genes are present in the sample and the relative levels of expression.
20 Signal intensities are used to determine relative amounts of sequences in the sample. Computer implemented design methods optimize the choice of the target subsequences.

A second specific embodiment of the invention, termed colony calling ("CC"), gathers only target subsequence
25 presence information for all target subsequences for arrayed, individual single sequence clones in a library, with cDNA libraries being preferred. The target subsequences are carefully chosen according to computer implemented design methods of this invention to have a maximum information
30 content and to be minimum in number. Preferably from 10-20 subsequences are sufficient to characterize the expressed cDNA in a tissue. In order to increase the specificity and reliability of hybridization to the typically short DNA subsequences, preferable recognition means are PNAs.
35 Degenerate sets of longer DNA oligomers having a common, short, shared, targ t sequence can also be used as a recognition means. A computer implement d gene calling step

uses th pattern of hits in conjunction with a database of
DNA sequences to determine which genes are present in the
sample and the relative levels of expression.

The embodiments of this invention preferably
5 generate measurements that are precise, reproducible, and
free of noise. Measurement noise in QEA™ is typically
created by generation or amplification of unwanted DNA
fragments, and special steps are preferably taken to avoid
any such unwanted fragments. Measurement noise in colony
10 calling is typically created by mis-hybridization of probes,
or recognition means, to colonies. High stringency reaction
conditions and DNA mimics with increased hybridization
specificity may be used to minimize this noise. DNA mimics
are polymers composed of subunits capable of specific,
15 Watson-Crick-like hybridization with DNA. Also useful to
minimize noise in colony calling are improved hybridization
detection methods. Instead of the conventional detection
methods based on probe labeling with fluorochromes, new
methods are based on light scattering by small 100-200 um
20 particles that are aggregated upon probe hybridization
(Stimson et al., 1995, "Real-time detection of DNA
hybridization and melting on oligonucleotide arrays by using
optical wave guides", Proc. Natl. Acad. Sci. USA 92:6379-
6383). In this method, the hybridization surface forms one
25 surface of a light pipe or optical wave guide, and the
scattering induced by these aggregated particles causes light
to leak from the light pipe. In this manner hybridization is
revealed as an illuminated spot of leaking light on a dark
background. This latter method makes hybridization detection
30 more rapid by eliminating the need for a washing step between
the hybridization and detection steps. Further by using
variously sized and shaped particles with different light
scattering properties, multiple probe hybridizations can be
detected from one colony.
35   Further, the embodiments of the invention can be
adapted to aut mation by eliminating non-automatabl steps,
such as extractions or buffer exchang s. The embodim nts of

- 11 -

th invention facilitate efficient analysis by permitting
multipl recognition m ans to be tested in on r action and
by utilizing multiple, distinguishable labeling of the
recognition means, so that signals may be simultaneously
5 detected and measured. Preferably, for QEA™ embodiments,
this labeling is by multiple fluorochromes. For the CC
embodiments, detection is preferably done by the light
scattering methods with variously sized and shaped particles.
       An increase in sensitivity as well as an increase
10 in the number of resolvable fluorescent labels can be
achieved by the use of fluorescent, energy transfer, dye-
labeled primers. Other detection methods, preferable when
the genes being identified will be physically isolated from
the gel for later sequencing or use as experimental probes,
15 include the use of silver staining gels or of radioactive
labeling. Since these methods do not allow for multiple
samples to be run in a single lane, they are less preferable
when high throughput is needed.
       Because this invention achieves rapid and
20 economical determination of quantitative gene expression in
tissue or other samples, it has considerable medical and
research utility. In medicine, as more and more diseases are
recognized to have important genetic components to their
etiology and development, it is becoming increasingly useful
25 to be able to assay the genetic makeup and expression of a
tissue sample. For example, the presence and expression of
certain genes or their particular alleles are prognostic or
risk factors for disease (including disorders). Several
examples of such diseases are found among the
30 neurodegenerative diseases, such as Huntington's disease and
ataxia-telangiectasia. Several cancers, such as
neuroblastoma, can now be linked to specific genetic defects.
Finally, gene expression can also determine the presence and
classification of those foreign pathogens that are difficult
35 or impossible to culture in vitro but which nevertheless
expr ss their own unique g nes.

Disease progression is reflected in changes in genetic expression of an affected tissue. For example, expression of particular tumor promoter genes and lack of expression of particular tumor suppressor genes is now known
5 to correlate with the progression of certain tumors from normal tissue, to hyperplasia, to cancer *in situ*, and to metastatic cancer. Return of a cell population to a normal pattern of gene expression, such as by using anti-sense technology, can correlate with tumor regression. Therefore,
10 knowledge of gene expression in a cancerous tissue can assist in staging and classifying this disease.

Expression information can also be used to chose and guide therapy. Accurate disease classification and staging or grading using gene expression information can
15 assist in choosing initial therapies that are increasingly more precisely tailored to the precise disease process occurring in the particular patient. Gene expression information can then track disease progression or regression, and such information can assist in monitoring the success or
20 changing the course of an initial therapy. A therapy is favored that results in a regression towards normal of an abnormal pattern of gene expression in an individual, while therapy which has little effect on gene expression or its progression can need modification. Such monitoring is now
25 useful for cancers and will become useful for an increasing number of other diseases, such as diabetes and obesity. Finally, in the case of direct gene therapy, expression analysis directly monitors the success of treatment.

In biological research, rapid and economical assay
30 for gene expression in tissue or other samples has numerous applications. Such applications include, but are not limited to, for example, in pathology examining tissue specific genetic response to disease, in embryology determining developmental changes in gene expression, in pharmacology
35 assessing direct and indirect effects of drugs on gen expression. In thes applications, this invention can be applied, .g., t *in vitro* c 11 populations or cell lines, to

- 13 -

in vivo animal models of disease or other processes, to human
sampl s, to purified cell populations perhaps drawn from
actual wild-type occurrences, and to tissue samples
containing mixed cell populations. The cell or tissue
5 sources can advantageously be a plant, a single celled
animal, a multicellular animal, a bacterium, a virus, a
fungus, or a yeast, etc. The animal can advantageously be
laboratory animals used in research, such as mice engineered
or bread to have certain genomes or disease conditions or
10 tendencies. The *in vitro* cell populations or cell lines can
be exposed to various exogenous factors to determine the
effect of such factors on gene expression. Further, since an
unknown signal pattern is indicative of an as yet unknown
gene, this invention has important use for the discovery of
15 new genes. In medical research, by way of further example,
use of the methods of this invention allow correlating gene
expression with the presence and progress of a disease and
thereby provide new methods of diagnosis and new avenues of
therapy which seek to directly alter gene expression.
20   This invention includes various embodiments and
aspects, several of which are described below.

         In a first embodiment, the invention provides a
method for identifying, classifying, or quantifying one or
more nucleic acids in a sample comprising a plurality of
25 nucleic acids having different nucleotide sequences, said
method comprising probing said sample with one or more
recognition means, each recognition means recognizing a
different target nucleotide subsequence or a different set of
target nucleotide subsequences; generating one or more
30 signals from said sample probed by said recognition means,
each generated signal arising from a nucleic acid in said
sample and comprising a representation of (i) the length
between occurrences of target subsequences in said nucleic
acid and (ii) the identities of said target subsequences in
35 said nucleic acid or the identities of said sets of target
subs quences among which is included the target subsequences
in said nucleic acid; and searching a nucleotide sequence

database to determin sequences that match or the absence of
any sequences that match said on or more generated signals,
said database comprising a plurality of known nucleotide
sequences of nucleic acids that may be present in the sample,
5 a sequence from said database matching a generated signal
when the sequence from said database has both (i) the same
length between occurrences of target subsequences as is
represented by the generated signal and (ii) the same target
subsequences as is represented by the generated signal, or
10 target subsequences that are members of the same sets of
target subsequences represented by the generated signal,
whereby said one or more nucleic acids in said sample are
identified, classified, or quantified.

     This invention further provides in the first
15 embodiment additional methods wherein each recognition means
recognizes one target subsequence, and wherein a sequence
from said database matches a generated signal when the
sequence from said database has both the same length between
occurrences of target subsequences as is represented by the
20 generated signal and the same target subsequences as
represented by the generated signal, or optionally wherein
each recognition means recognizes a set of target
subsequences, and wherein a sequence from said database
matches a generated signal when the sequence from said
25 database has both the same length between occurrences of
target subsequences as is represented by the generated
signal, and target subsequences that are members of the sets
of target subsequences represented by the generated signal.

     This invention further provides in the first
30 embodiment additional methods further comprising dividing
said sample of nucleic acids into a plurality of portions and
performing the methods of this object individually on a
plurality of said portions, wherein a different one or more
recognition means are used with each portion.

35   This invention further provides in the first
emb diment additional methods wherein the quantitative
abundance of a nucleic acid comprising a particular

- 15 -

nucleotide sequ nce in th sampl is determin d from the
quantitative level of the one or more signals generated by
said nucleic acid that are determined to match said
particular nucleotide sequence.

5   This invention further provides in the first
embodiment additional methods wherein said plurality of
nucleic acids are DNA, and optionally wherein the DNA is
cDNA, and optionally wherein the cDNA is prepared from a
plant, an single celled animal, a multicellular animal, a
10 bacterium, a virus, a fungus, or a yeast, and optionally
wherein the cDNA is of total cellular RNA or total cellular
poly(A) RNA.

          This invention further provides in the first
embodiment additional methods wherein said database comprises
15 substantially all the known expressed sequences of said
plant, single celled animal, multicellular animal, bacterium,
or yeast.

          This invention further provides in the first
embodiment additional methods wherein the recognition means
20 are one or more restriction endonucleases whose recognition
sites are said target subsequences, and wherein the step of
probing comprises digesting said sample with said one or more
restriction endonucleases into fragments and ligating double
stranded adapter DNA molecules to said fragments to produce
25 ligated fragments, each said adapter DNA molecule comprising
(i) a shorter stand having no 5' terminal phosphates and
consisting of a first and second portion, said first portion
at the 5' end of the shorter strand being complementary to
the overhang produced by one of said restriction
30 endonucleases and (ii) a longer strand having a 3' end
subsequence complementary to said second portion of the
shorter strand; and wherein the step of generating further
comprises melting the shorter strand from the ligated
fragments, contacting the sample with a DNA polymerase,
35 extending the ligated fragments by synthesis with the DNA
polymerase to produce blunt-ended double strand d DNA
fragments, and amplifying the blunt-ended fragments by a

method comprising contacting said blunt-ended fragments with
a DNA polymerase and prim r oligodeoxynucl otides, said
primer oligodeoxynucleotides comprising the longer adapter
strand, and said contacting being at a temperature not
5 greater than the melting temperature of the primer
oligodeoxynucleotide from a strand of the blunt-ended
fragments complementary to the primer oligodeoxynucleotide
and not less than the melting temperature of the shorter
strand of the adapter nucleic acid from the blunt-ended
10 fragments.

This invention further provides in the first
embodiment additional methods wherein the recognition means
are one or more restriction endonucleases whose recognition
sites are said target subsequences, and wherein the step of
15 probing further comprises digesting the sample with said one
or more restriction endonucleases.

This invention further provides in the first
embodiment additional methods further comprising identifying
a fragment of a nucleic acid in the sample which generates
20 said one or more signals; and recovering said fragment, and
optionally wherein the signals generated by said recovered
fragment do not match a sequence in said nucleotide sequence
database, and optionally further comprising using at least a
hybridizable portion of said fragment as a hybridization
25 probe to bind to a nucleic acid that can generate said
fragment upon digestion by said one or more restriction
endonucleases.

This invention further provides in the first
embodiment additional methods wherein the step of generating
30 further comprises after said digesting removing from the
sample both nucleic acids which have not been digested and
nucleic acid fragments resulting from digestion at only a
single terminus of the fragments, and optionally wherein
prior to digesting, the nucleic acids in the sample are each
35 bound at one terminus to a biotin molecule or to a hapten
molecule, and said r moving is carried out by a method which
comprises contacting the nucleic acids in the sample with

streptavidin or avidin or with an anti-hapten antibody, respectively, affixed to a solid support.

This invention further provides in the first embodiment additional methods wherein said digesting with
5 said one or more restriction endonucleases leaves single-stranded nucleotide overhangs on the digested ends.

This invention further provides in the first embodiment additional methods wherein the step of probing further comprises hybridizing double-stranded adapter nucleic
10 acids with the digested sample fragments, each said adapter nucleic acid having an end complementary to said overhang generated by a particular one of the one or more restriction endonucleases, and ligating with a ligase a strand of said adapter nucleic acids to the 5' end of a strand of the
15 digested sample fragments to form ligated nucleic acid fragments.

This invention further provides in the first embodiment additional methods wherein said digesting with said one or more restriction endonucleases and said ligating
20 are carried out in the same reaction medium, and optionally wherein said digesting and said ligating comprises incubating said reaction medium at a first temperature and then at a second temperature, in which said one or more restriction endonucleases are more active at the first temperature than
25 the second temperature and said ligase is more active at the second temperature that the first temperature, or wherein said incubating at said first temperature and said incubating at said second temperature are performed repetitively.

This invention further provides in the first
30 embodiment additional methods wherein the step of probing further comprises prior to said digesting removing terminal phosphates from DNA in said sample by incubation with an alkaline phosphatase, and optionally wherein said alkaline phosphatase is heat labile and is heat inactivated prior to
35 said digesting.

This invention further provides in the first embodiment additional methods wherein said generating step

compris s amplifying the ligat d nucleic acid fragments, and
optionally wherein said amplifying is carried out by use of a
nucleic acid polymerase and primer nucleic acid strands, said
primer nucleic acid strands being capable of priming nucleic
5 acid synthesis by said polymerase, and optionally wherein the
primer nucleic acid strands have a G+C content of between 40%
and 60%.

    This invention further provides in the first
embodiment additional methods wherein each said adapter
10 nucleic acid has a shorter strand and a longer strand, the
longer strand being ligated to the digested sample fragments,
and said generating step comprises prior to said amplifying
step the melting of the shorter strand from the ligated
fragments, contacting the ligated fragments with a DNA
15 polymerase, extending the ligated fragments by synthesis with
the DNA polymerase to produce blunt-ended double stranded DNA
fragments, and wherein the primer nucleic acid strands
comprise a hybridizable portion the sequence of said longer
strands, or optionally comprise the sequence of said longer
20 strands, each different primer nucleic acid strand priming
amplification only of blunt ended double stranded DNA
fragments that are produced after digestion by a particular
restriction endonuclease.

    This invention further provides in the first
25 embodiment additional methods wherein each primer nucleic
acid strand is specific for a particular restriction
endonuclease, and further comprises at the 3' end of and
contiguous with the longer strand sequence the portion of the
restriction endonuclease recognition site remaining on a
30 nucleic acid fragment terminus after digestion by the
restriction endonuclease, or optionally wherein each said
primer specific for a particular restriction endonuclease
further comprises at its 3' end one or more nucleotides 3' to
and contiguous with the remaining portion of the restriction
35 endonuclease recognition site, whereby the ligated nucleic
acid fragment amplified is that comprising said remaining
portion of said restriction endonuclease recognition site

contiguous to said one or more additional nucleotides, and
optionally such that said primers comprising a particular
said one or more additional nucleotides can be
distinguishably detected from said primers comprising a
5 different said one or more additional nucleotides.

This invention further provides in the first
embodiment additional methods wherein during said amplifying
step the primer nucleic acid strands are annealed to the
ligated nucleic acid fragments at a temperature that is less
10 than the melting temperature of the primer nucleic acid
strands from strands complementary to the primer nucleic acid
strands but greater than the melting temperature of the
shorter adapter strands from the blunt-ended fragments.

This invention further provides in the first
15 embodiment additional methods wherein the recognition means
are oligomers of nucleotides, nucleotide-mimics, or a
combination of nucleotides and nucleotide-mimics, which are
specifically hybridizable with the target subsequences, and
optionally further provides additional methods wherein the
20 step of generating comprises amplifying with a nucleic acid
polymerase and with primers comprising said oligomers,
whereby fragments of nucleic acids in the sample between
hybridized oligomers are amplified.

This invention further provides in the first
25 embodiment additional methods wherein said signals further
comprise a representation of whether an additional target
subsequence is present on said nucleic acid in the sample
between said occurrences of target subsequences, and
optionally wherein said additional target subsequence is
30 recognized by a method comprising contacting nucleic acids in
the sample with oligomers of nucleotides, nucleotide-mimics,
or mixed nucleotides and nucleotide-mimics, which are
hybridizable with said additional target subsequence.

This invention further provides in the first
35 embodiment additional methods wherein the step of generating
comprises suppressing said signals when an additional target
subsequence is present on said nucleic acid in the sample

between said occurrences of target subsequences, and
optionally wherein, when the step of generating comprises
amplifying nucleic acids in the sample, said additional
target subsequence is recognized by a method comprising
5 contacting nucleic acids in the sample with (a) oligomers of
nucleotides, nucleotide-mimics, or mixed nucleotides and
nucleotide-mimics, which hybridize with said additional
target subsequence and disrupt the amplifying step; or (b)
restriction endonucleases which have said additional target
10 subsequence as a recognition site and digest the nucleic
acids in the sample at the recognition site.

This invention further provides in the first
embodiment additional methods wherein the step of generating
further comprises separating nucleic acid fragments by
15 length, and optionally wherein the step of generating further
comprises detecting said separated nucleic acid fragments,
and optionally wherein said detecting is carried out by a
method comprising staining said fragments with silver,
labeling said fragments with a DNA intercalating dye, or
20 detecting light emission from a fluorochrome label on said
fragments.

This invention further provides in the first
embodiment additional methods wherein said representation of
the length between occurrences of target subsequences is the
25 length of fragments determined by said separating and
detecting steps.

This invention further provides in the first
embodiment additional methods wherein said separating is
carried out by use of liquid chromatography, mass
30 spectrometry, or electrophoresis, and optionally wherein said
electrophoresis is carried out in a slab gel or capillary
configuration using a denaturing or non-denaturing medium.

This invention further provides in the first
embodiment additional methods wherein a predetermined one or
35 more nucleotide sequences in said database are of interest,
and wherein the target subsequences are such that said
sequences of inter st generate at least one signal that is

- 21 -

not generated by any other sequence likely to be present in the sample, and optionally wherein the nucleotide sequences of interest are a majority of sequences in said database.

This invention further provides in the first
5 embodiment additional methods wherein the target subsequences have a probability of occurrence in the nucleotide sequences in said database of from approximately 0.01 to approximately 0.30.

This invention further provides in the first
10 embodiment additional methods wherein the target subsequences are such that the majority of sequences in said database contain on average a sufficient number of occurrences of target subsequences in order to on average generate a signal that is not generated by any other nucleotide sequence in
15 said database, and optionally wherein the number of pairs of target subsequences present on average in the majority of sequences in said database is no less than 3, and wherein the average number of signals generated from the sequences in said database is such that the average difference between
20 lengths represented by the generated signals is greater than or equal to 1 base pair.

This invention further provides in the first embodiment additional methods wherein the target subsequences have a probability of occurrence, p, approximately given by
25 the solution of

$$\frac{R(R + 1)p^2}{2} = A$$

and
30

$$\frac{L}{Np^2} = B$$

wherein N = the number of different nucleotide sequences in
35 said database; L = the average length of said different nucleotide sequences in said database; R = the numb r of recognition means; A = the number of pairs of target

subsequences present on average in said different nucl otide
sequences in said database; and B = the average difference
between lengths represented by the signals generated from the
nucleic acids in the sample, and optionally wherein A is
5 greater than or equal to 3 and wherein B is greater than or
equal to 1.

This invention further provides in the first
embodiment additional methods wherein the target subsequences
are selected according to the further steps comprising
10 determining a pattern of signals that can be generated and
the sequences capable of generating each such signal by
simulating the steps of probing and generating applied to
each sequences in said database of nucleotide sequences;
ascertaining the value of said determined pattern according
15 to an information measure; and choosing the target
subsequences in order to generate a new pattern that
optimizes the information measure, and optionally wherein
said choosing step selects target subsequences which comprise
the recognition sites of the one or more restriction
20 endonucleases, and optionally wherein said choosing step
selects target subsequences which comprise the recognition
sites of the one or more restriction endonucleases contiguous
with one or more additional nucleotides.

This invention further provides in the first
25 embodiment additional methods wherein a predetermined one or
more of the nucleotide sequences present in said database of
nucleotide sequences are of interest, and the information
measure optimized is the number of such said sequences of
interest which generate at least one signal that is not
30 generated by any other nucleotide sequence present in said
database, and optionally wherein said nucleotide sequences of
interest are a majority of the nucleotide sequences present
in said database.

This invention further provides in the first
35 embodiment additional methods wherein said choosing step is
by exhaustiv s arch of all combinations of target
subsequences of length less than approximately 10, or wherein

- 23 -

said step of ch osing target subsequences is by a method
comprising simulated annealing.

This invention further provides in the first
embodiment additional methods wherein the step of searching
5 further comprises determining a pattern of signals that can
be generated and the sequences capable of generating each
such signal by simulating the steps of probing and generating
applied to each sequence in said database of nucleotide
sequences; and finding the one or more nucleotide sequences
10 in said database that are able to generate said one or more
generated signals by finding in said pattern those signals
that comprise a representation of the (i) the same lengths
between occurrences of target subsequences as is represented
by the generated signal and (ii) the same target subsequences
15 as is represented by the generated signal, or target
subsequences that are members of the same sets of target
subsequences represented by the generated signal.

This invention further provides in the first
embodiment additional methods wherein the step of determining
20 further comprises searching for occurrences of said target
subsequences or sets of target subsequences in nucleotide
sequences in said database of nucleotide sequences; finding
the lengths between occurrences of said target subsequences
or sets of target subsequences in the nucleotide sequences of
25 said database; and forming the pattern of signals that can be
generated from the sequences of said database in which the
target subsequences were found to occur.

This invention further provides in the first
embodiment additional methods wherein said restriction
30 endonucleases generate 5' overhangs at the terminus of
digested fragments and wherein each double stranded adapter
nucleic acid comprises a shorter nucleic acid strand
consisting of a first and second contiguous portion, said
first portion being a 5' end subsequence complementary to the
35 overhang produced by one of said restriction endonucleases;
and a longer nucleic acid strand having a 3' nd subsequence
complementary to said second portion of the short r strand.

This inv nti n further provides in the first
embodiment additional methods wherein said shorter strand has
a melting temperature from a complementary strand of less
than approximately 68°C, and has no terminal phosphate, and
5 optionally wherein said shorter strand is approximately 12
nucleotides long.

This invention further provides in the first
embodiment additional methods wherein said longer strand has
a melting temperature from a complementary strand of greater
10 than approximately 68°C, is not complementary to any
nucleotide sequence in said database, and has no terminal
phosphate, and optionally wherein said ligated nucleic acid
fragments do not contain a recognition site for any of said
restriction endonucleases, and optionally wherein said longer
15 strand is approximately 24 nucleotides long and has a G+C
content between 40% and 60%.

This invention further provides in the first
embodiment additional methods wherein said one or more
restriction endonucleases are heat inactivated before said
20 ligating.

This invention further provides in the first
embodiment additional methods wherein said restriction
endonucleases generate 3' overhangs at the terminus of the
digested fragments and wherein each double stranded adapter
25 nucleic acid comprises a longer nucleic acid strand
consisting of a first and second contiguous portion, said
first portion being a 3' end subsequence complementary to the
overhang produced by one of said restriction endonucleases;
and a shorter nucleic acid strand complementary to the 3' end
30 of said second portion of the longer nucleic acid stand.

This invention further provides in the first
embodiment additional methods wherein said shorter strand has
a melting temperature from said longer strand of less than
approximately 68°C, and has no terminal phosphates, and
35 optionally wherein said shorter strand is 12 base pairs long.

This invention further provides in the first
embodiment additional m thods wherein said longer strand has

a m lting temp rature from a complementary strand of greater
than approximat ly 68°C, is not complementary to any
nucleotide sequence in said database, has no terminal
phosphate, and wherein said ligated nucleic acid fragments do
5 not contain a recognition site for any of said restriction
endonucleases, and optionally wherein said longer strand is
24 base pairs long and has a G+C content between 40% and 60%.

In a second embodiment, the invention provides a
method for identifying or classifying a nucleic acid
10 comprising probing said nucleic acid with a plurality of
recognition means, each recognition means recognizing a
target nucleotide subsequence or a set of target nucleotide
subsequences, in order to generate a set of signals, each
signal representing whether said target subsequence or one of
15 said set of target subsequences is present or absent in said
nucleic acid; and searching a nucleotide sequence database,
said database comprising a plurality of known nucleotide
sequences of nucleic acids that may be present in the sample,
for sequences matching said generated set of signals, a
20 sequence from said database matching a set of signals when
the sequence from said database (i) comprises the same target
subsequences as are represented as present, or comprises
target subsequences that are members of the sets of target
subsequences represented as present by the generated sets of
25 signals and (ii) does not comprise the target subsequences
represented as absent or that are members of the sets of
target subsequences represented as absent by the generated
sets of signals, whereby the nucleic acid is identified or
classified, and optionally wherein the set of signals are
30 represented by a hash code which is a binary number.

This invention further provides in the second
embodiment additional methods wherein the step of probing
generates quantitative signals of the numbers of occurrences
of said target subsequences or of members of said set of
35 target subsequences in said nucleic acid, and optionally
wher in a sequ nce matches said generated set of signals when
the sequence from said database comprises the same target

subsequences with the same numb r of occurrences in said
sequenc as in the quantitative signals and does not comprise
the target subsequences represented as absent or target
subsequences within the sets of target subsequences
5 represented as absent.

This invention further provides in the second
embodiment additional methods wherein said plurality of
nucleic acids are DNA.

This invention further provides in the second
10 embodiment additional methods wherein the recognition means
are detectably labeled oligomers of nucleotides, nucleotide-
mimics, or combinations of nucleotides and nucleotide-mimics,
and the step of probing comprises hybridizing said nucleic
acid with said oligomers, and optionally wherein said
15 detectably labeled oligomers are detected by a method
comprising detecting light emission from a fluorochrome label
on said oligomers or arranging said labeled oligomers to
cause light to scatter from a light pipe and detecting said
scattering, and optionally wherein the recognition means are
20 oligomers of peptido-nucleic acids, and optionally wherein
the recognition means are DNA oligomers, DNA oligomers
comprising universal nucleotides, or sets of partially
degenerate DNA oligomers.

This invention further provides in the second
25 embodiment additional methods wherein the step of searching
further comprises determining a pattern of sets of signals of
the presence or absence of said target subsequences or said
sets of target subsequences that can be generated and the
sequences capable of generating each set of signals in said
30 pattern by simulating the step of probing as applied to each
sequence in said database of nucleotide sequences; and
finding one or more nucleotide sequences that are capable of
generating said generated set of signals by finding in said
pattern those sets that match said generated set, where a set
35 of signals from said pattern matches a generated set of
signals when th s t from said patt rn (i) represents as
present the same target subsequences as are represented as

present or target subsequences that are members of the sets
of target subs quenc s repres nted as pr sent by the
generated sets of signals and (ii) represents as absent the
target subsequences represented as absent or that are members
5 of the sets of target subsequences represented as absent by
the generated sets of signals.

This invention further provides in the second
embodiment additional methods wherein the target subsequences
are selected according to the further steps comprising
10 determining (i) a pattern of sets of signals representing the
presence or absence of said target subsequences or of said
sets of target subsequences that can be generated, and (ii)
the sequences capable of generating each set of signals in
said pattern by simulating the step of probing as applied to
15 each sequence in said database of nucleotide sequences;
ascertaining the value of said pattern generated according to
an information measure; and choosing the target subsequences
in order to generate a new pattern that optimizes the
information measure.

20   This invention further provides in the second
embodiment additional methods wherein the information measure
is the number of sets of signals in the pattern which are
capable of being generated by one or more sequences in said
database, or optionally wherein the information measure is
25 the number of sets of signals in the pattern which are
capable of being generated by only one sequence in said
database.

This invention further provides in the second
embodiment additional methods wherein said choosing step is
30 by a method comprising exhaustive search of all combination
of target subsequences of length less than approximately 10,
or optionally wherein said choosing step is by a method
comprising simulated annealing.

This invention further provides in the second
35 embodiment additional methods wherein the step of determining
by simulating further compris s searching for the presence or
absence of said target subsequ nces or sets of target

subs quences in each nucleotide sequence in said database of
nucleotide sequences; and forming the pattern of s ts of
signals that can be generated from said sequences in said
database, and optionally where the step of searching is
5 carried out by a string search, and optionally wherein the
step of searching comprises counting the number of
occurrences of said target subsequences in each nucleotide
sequence.

This invention further provides in the second
10 embodiment additional methods wherein the target subsequences
have a probability of occurrence in a nucleotide sequence in
said database of nucleotide sequences of from 0.01 to 0.6, or
optionally wherein the target subsequences are such that the
presence of one target subsequence in a nucleotide sequence
15 in said database of nucleotide sequences is substantially
independent of the presence of any other target subsequence
in the nucleotide sequence, or optionally wherein fewer than
approximately 50 target subsequences are selected.

In a third embodiment, the invention provides a
20 programmable apparatus for analyzing signals comprising an
inputting device for inputting one or more actual signals
generated by probing a sample comprising a plurality of
nucleic acids with recognition means, each recognition means
recognizing a target nucleotide subsequence or a set of
25 target nucleotide subsequences, said signals comprising a
representation of (i) the length between occurrences of said
target subsequences in a nucleic acid of said sample, and
(ii) the identities of said target subsequences in said
nucleic acid, or the identities of said sets of target
30 subsequences among which is included the target subsequences
in said nucleic acid; a searching device operatively coupled
to said accepting device for searching a sequence in a
nucleotide sequence database for occurrences of said target
subsequences or target subsequences that are members of said
35 sets of target subsequences, and for the length between such
occurrences, said database comprising a plurality of known
nucleotid sequences that may be present in said sample; a

- 29 -

comparing device operatively coupled t said accepting device
and to said s arching device for finding a match between said
one or more actual signals and a sequence in said database,
said one or more actual signals matching a sequence from said

5 database when the sequence from said database has both (i)
the same length between occurrences of target subsequences as
is represented by said one or more actual signals and (ii)
the same target subsequences as is represented by said one or
more actual signals or target subsequences that are members

10 of the same sets of target subsequences represented by said
one or more actual signals; and a control device operatively
coupled to said comparing device for causing said comparing
to be done for sequences in the database and for outputting
those database sequences that match said one or more actual

15 signals, and optionally wherein said searching device
searches for said target subsequences or a set of target
nucleotide subsequences in said database sequences by
performing a string comparison of the nucleotides in said
subsequences with those in said database sequence.

20   This invention further provides in the third
embodiment that said control device further comprises causing
said searching device to search substantially all sequences
in said database in order to determine a pattern of signals
that can be generated by probing said sample with said

25 recognition means, and wherein said control device further
causes said comparing device to find any matches between said
one or more actual signals and said pattern of signals, said
one or more actual signals matching a signal in said pattern
of signals when the signal from said pattern represents (i)

30 the same length between occurrences of target subsequences as
is represented by said one or more actual signals and (ii)
the same target subsequences as is represented by said one or
more actual signals or target subsequences that are members
of the same sets of target subsequences represented by said

35 one or more actual signals.
            This invention further provid s in the third
mbodim nt that said sample of nucleic acids comprises cDNA

- 30 -

from RNA of a cell or tissu type,and said database comprises
DNA s quences that are likely to be xpressed by d cell or
tissue type.

This invention further provides in the third
5 embodiment a computer readable memory that can be used to
direct a programmable apparatus to function for analyzing
signals according to steps comprising inputting one or more
actual signals generated by probing a sample comprising a
plurality of nucleic acids with recognition means, each
10 recognition means recognizing a target nucleotide subsequence
or a set of target nucleotide subsequences, said signals
comprising a representation of (i) the length between
occurrences of said target subsequences in a nucleic acid of
said sample, and (ii) the identities of said target
15 subsequences in said nucleic acid, or the identities of said
sets of target subsequences among which is included the
target subsequences in said nucleic acid; searching a
sequence in a nucleotide sequence database for occurrences of
said target subsequences or target subsequences that are
20 members of said sets of target subsequences, and for the
length between such occurrences, said database comprising a
plurality of known nucleotide sequences that may be present
in said sample; matching said one or more actual signals and
a sequence in said database when the sequence in said
25 database has both (i) the same length between occurrences of
target subsequences as is represented by said one or more
actual signals and (ii) the same target subsequences as is
represented by said one or more actual signals, or target
subsequences that are members of the same sets of target
30 subsequences as is represented by said one or more actual
signals; and repetitively performing said searching and
matching steps for the majority of sequences in the database
and outputting those database sequences that match said one
or more actual signals, or alternatively a computer readable
35 memory for directing a programmable apparatus to function in
the manner of the third object.

In a fourth mbodim nt, the invention provides a
programmable apparatus for selecting target subsequences
comprising an initial selection device for selecting initial
target subsequences or initial sets of target subsequences; a
5 first control device; a search device operatively coupled to
said initial selection device and to said first control
device (i) for searching sequences in a nucleotide sequence
database for occurrences of said initial target subsequences
or occurrences of target subsequences that are members of
10 said initial sets of target subsequences and for the length
between such occurrences and (ii) for determining an initial
pattern of signals that can be generated from said selected
initial target subsequences or said initial sets of target
subsequences, said database comprising a plurality of known
15 nucleotide sequences, said signals comprising a
representation of (i) the length between said occurrences in
a sequence in said database, and (ii) the identities of said
initial target subsequences that occur in said sequence in
said database, or the identities of target subsequences that
20 are members of the same initial sets of target subsequences
that occur in said sequence in said database; and an
ascertaining device operatively coupled to said searching
device and to said first control device for ascertaining the
value of said determined initial pattern according to an
25 information measure; and wherein said first control device
causes further target subsequences to be selected and causes
the search device to determine a further pattern of signals
and the ascertaining device to ascertain a further value of
said information measure and accepts the further target
30 subsequences when said further pattern optimizes said further
value of said information measure.

This invention further provides in the fourth
object that a predetermined one or more of the sequences in
said database are of interest, and wherein said ascertaining
35 device ascertains the value of an information measure by
counting the numb r of such sequences of inter st which
generate in said determined pattern at least ne signal that

is not generated by any other sequence in said database, and optionally that said one or more of th sequences of interest comprise substantially all the sequences in said database.

This invention further provides in the fourth
5 embodiment that said first control device optimizes the value of said information measure according to a method of exhaustive search, wherein said first control device selects further target subsequences of length less than approximately 10 and accepts the further target subsequences if said
10 further value of said information measure is greater than the previous value.

This invention further provides in the fourth embodiment that said first control device optimizes the value of said information measure according to a method comprising
15 simulated annealing, wherein said first control device repeatedly selects further target subsequences and accepts the further target subsequences if said further value of said information measure is not decreased by greater than a probabilistic factor dependent on a simulated-temperature,
20 and wherein said programmable apparatus further comprises a second control device operatively coupled to said first control device for decreasing said simulated-temperature as said first control device selects further target subsequences, and optionally wherein said probabilistic
25 factor is an exponential function of the negative of the decrease in the information measure divided by said simulated-temperature.

This invention further provides in the fourth embodiment that the database comprises a majority of known
30 DNA sequences that are likely to be expressed by one or more cell types.

This invention further provides in the fourth embodiment a computer readable memory that can be used to direct a programmable apparatus to function for selecting·
35 target subsequences according to steps comprising selecting initial targ t subsequences or initial sets of target subsequences; s arching a sequenc in a nucleotide sequence

database f r occurr nc s of said initial targ t subsequences
or occurrences of target subsequ nces that are members of
said initial sets of target subsequences and for the length
between such occurrences, said database comprising a
5 plurality of known nucleotide sequences that may be present
in said sample; determining an initial pattern of signals
that can be generated from said selected initial target
subsequences or said initial sets of target subsequences,
said signals comprising a representation of (i) the length
10 between said occurrences in a sequence in said database, and
(ii) the identities of said initial target subsequences that
occur in said sequence in said database, or the identities of
target subsequences that are members of the initial sets of
target subsequences that occur in said sequence in said
15 database; ascertaining the value of said determined initial
pattern according to an information measure; and repetitively
performing said selecting, searching, determining, and
ascertaining steps to determine a further pattern of signals
and a further value of said information measure, and
20 accepting the further target subsequences when said further
pattern optimizes said further value of said information
measure, or alternatively a computer readable memory for
directing a programmable apparatus to function in the manner
of the fourth object.
25   In a fifth embodiment, the invention provides a
programmable apparatus for displaying data comprising a
selecting device for selecting target subsequences or sets of
target subsequences, such that recognition means for
recognizing said target subsequences or said sets of target
30 subsequences can be used to generate signals by probing a
sample comprising a plurality of nucleic acids, said signals
comprising a representation of (i) the length between
occurrences of said target subsequences in a nucleic acid of
said sample and (ii) the identities of said target
35 subsequences in said nucleic acid or the identities of said
sets of target subsequences among which are included the
target subsequences in said nucleic acid; an inputting device

for inputting one or more actual signals generated by probing
said sample with said recognition means; an analyzing device
for analyzing signals operatively coupled to said selecting
and inputting devices that determines which sequences in a
5 nucleotide sequence database can generate said actual signals
when subject to said recognition means, said database
comprising a plurality of known nucleotide sequences that may
be present in said sample; an input/output device operatively
coupled to said selecting, inputting, and analyzing devices
10 that inputs user requests and controls the selecting device
to select target subsequences or sets of target subsequences,
controls the inputting device to accept actual signals,
controls the analyzing device to find the sequences in said
database that can generate said actual signals, and displays
15 output comprising said actual signals and said sequences in
said database that can generate said actual signals.

    This invention further provides in the fifth
embodiment that said sample is a cDNA sample prepared from a
tissue specimen, and the apparatus further comprises a
20 storage device operatively coupled to the input/output device
for storing indications of the origin of said tissue specimen
and information concerning said tissue specimen, and wherein
said indications can be displayed upon user input, and
optionally that the indications and information concerning
25 said tissue specimen comprises histological information
comprising tissue images.

    This invention further provides in the fifth
embodiment additional apparatus further comprising one or
more instrument devices for probing said sample with said
30 recognition means and for generating said actual signals; and
a control device operatively coupled to said one or more
instrument devices and to said input/output device for
controlling the operation of said instrument devices, wherein
said user can input control commands for control of said
35 instrument devices and receive output concerning the status
of said instrument devices, and optionally wherein one or
more of said selecting, inputting, analyzing, and

input/output devices are physically collocated with each
other, or are physically spaced apart from each other and are
connected by a communication medium for exchanges of commands
and information.

5   This invention further provides in the fifth
embodiment a computer readable memory that can be used to
direct a programmable apparatus to function for displaying
data according to steps comprising selecting target
subsequences or sets of target subsequences, such that

10 recognition means for recognizing said target subsequences or
said sets of target subsequences can be used to generate
signals by probing a sample comprising a plurality of nucleic
acids, said signals comprising a representation of (i) the
length between occurrences of said target subsequences in a

15 nucleic acid of said sample and (ii) the identities of said
target subsequences in said nucleic acid or the identities of
said sets of target subsequences among which are included the
target subsequences in said nucleic acid inputting one or
more actual signals generated by probing said sample with

20 said recognition means analyzing said one or more actual
signals to determine which sequences in a nucleotide sequence
database can generate said actual signals when subject to
said recognition means, said database comprising a plurality
of known nucleotide sequences that may be present in said

25 sample; and inputting user requests to control said selecting
step to select target subsequences or sets of target
subsequences, said inputting step to input actual signals,
and said analyzing step to find the sequences in said
database that can generate said actual signals, and

30 outputting in response to further user requests information
comprising said actual signals and said sequences in said
database that can generate said actual signals, or
alternatively a computer readable memory for directing a
programmable apparatus to function in the manner of the fifth

35 object.

        In a sixth emb diment, the inventi n provides a
meth d for identifying, classifying, or quantifying DNA

molecul s in a sample of DNA molecules having a plurality of
different nucleotide sequences, the method comprising the
steps of digesting said sample with one or more restriction
endonucleases, each said restriction endonuclease recognizing
5 a subsequence recognition site and digesting DNA at said
recognition site to produce fragments with 5' overhangs;
contacting said fragments with shorter and longer
oligodeoxynucleotides, each said shorter oligodeoxynucleotide
hybridizable with a said 5' overhang and having no terminal
10 phosphates, each said longer oligodeoxynucleotide
hybridizable with a said shorter oligodeoxynucleotide;
ligating said longer oligodeoxynucleotides to said 5'
overhangs on said DNA fragments to produce ligated DNA
fragments; extending said ligated DNA fragments by synthesis
15 with a DNA polymerase to produce blunt-ended double stranded
DNA fragments; amplifying said blunt-ended double stranded
DNA fragments by a method comprising contacting said DNA
fragments with a DNA polymerase and primer
oligodeoxynucleotides, each said primer oligodeoxynucleotide
20 having a sequence comprising that of one of the longer
oligodeoxynucleotides; determining the length of the
amplified DNA fragments; and searching a DNA sequence
database, said database comprising a plurality of known DNA
sequences that may be present in the sample, for sequences
25 matching one or more of said fragments of determined length,
a sequence from said database matching a fragment of
determined length when the sequence from said database
comprises recognition sites of said one or more restriction
endonucleases spaced apart by the determined length, whereby
30 DNA molecules in said sample are identified, classified, or
quantified.

            This invention further provides in the sixth
embodiment additional methods wherein the sequence of each
primer oligodeoxynucleotide further comprises 3' to and
35 contiguous with the sequence of the longer
oligodeoxynucleotide the p rtion f the recognition site of
said on or more restriction ndonucleases remaining on a DNA

fragment terminus after digestion, said r maining portion
being 5' to and contiguous with one or more additional
nucleotides, and wherein a sequence from said database
matches a fragment of determined length when the sequence
5 from said database comprises subsequences that are the
recognition sites of said one or more restriction
endonucleases contiguous with said one or more additional
nucleotides and when the subsequences are spaced apart by the
determined length.

10   This invention further provides in the sixth
embodiment additional methods wherein said determining step
further comprises detecting the amplified DNA fragments by a
method comprising staining said fragments with silver.

             This invention further provides in the sixth
15 embodiment additional methods wherein said
oligodeoxynucleotide primers are detectably labeled, wherein
the determining step further comprises detection of said
detectable labels, and wherein a sequence from said database
matches a fragment of determined length when the sequence
20 from said database comprises recognition sites of the one or
more restriction endonucleases, said recognition sites being
identified by the detectable labels of said
oligodeoxynucleotide primers, said recognition sites being
spaced apart by the determined length, and optionally wherein
25 said determining step further comprises detecting the
amplified DNA fragments by a method comprising labeling said
fragments with a DNA intercalating dye or detecting light
emission from a fluorochrome label on said fragments.

             This invention further provides in the sixth
30 embodiment additional steps further comprising, prior to said
determining step, the step of hybridizing the amplified DNA
fragments with a detectably labeled oligodeoxynucleotide
complementary to a subsequence, said subsequence differing
from said recognition sites of said one or more restriction
35 endonucleases, wherein the determining step further comprises
detecting said d tectable label of said oligode xynucleotide,
and wherein a s quence from said databas match s a fragm nt

of determined length when the sequence from said database
furth r comprises said subsequence between the recognition
sites of said one or more restriction endonucleases.

This invention further provides in the sixth
5 embodiment additional methods wherein the one or more
restriction endonucleases are pairs of restriction
endonucleases, the pairs being selected from the group
consisting of Acc56I and HindIII, Acc65I and NgoMI, BamHI and
EcoRI, BglII and HindIII, BglII and NgoMI, BsiWI and BspHI,
10 BspHI and BstYI, BspHI and NgoMI, BsrGI and EcoRI, EagI and
EcoRI, EagI and HindIII, EagI and NcoI, HindIII and NgoMI,.
NgoMI and NheI, NgoMI and SpeI, BglII and BspHI, Bsp120I and
NcoI, BssHII and NgoMI, EcoRI and HindIII, and NgoMI and
XbaI, or wherein the step of ligating is performed with T4
15 DNA ligase.

This invention further provides in the sixth
embodiment additional methods wherein the steps of digesting,
contacting, and ligating are performed simultaneously in the
same reaction vessel, or optionally wherein the steps of
20 digesting, contacting, ligating, extending, and amplifying
are performed in the same reaction vessel.

This invention further provides in the sixth
embodiment additional methods wherein the step of determining
the length is performed by electrophoresis.
25   This invention further provides in the sixth
embodiment additional methods wherein the step of searching
said DNA database further comprises determining a pattern of
fragments that can be generated and for each fragment in said
pattern those sequences in said DNA database that are capable
30 of generating the fragment by simulating the steps of
digesting with said one or more restriction endonucleases,
contacting, ligating, extending, amplifying, and determining
applied to each sequence in said DNA database; and finding
the sequences that are capable of generating said one or more
35 fragments of determined length by finding in said pattern one
or more fragments that hav the same length and recognition
sites as said one or mor fragments of determin d length..

This invention further provides in the sixth
embodim nt additional m thods wherein the st ps of digesting
and ligating go substantially to completion.

This invention further provides in the sixth
5 embodiment additional methods wherein the DNA sample is cDNA
prepared from mRNA, and optionally wherein the DNA is of RNA
from a tissue or a cell type derived from a plant, a single
celled animal, a multicellular animal, a bacterium, a virus,
a fungus, a yeast, or a mammal, and optionally wherein the
10 mammal is a human, and optionally wherein the mammal is a
human having or suspected of having a diseased condition, and
optionally wherein the diseased condition is a malignancy.

In a seventh embodiment, this invention provides
additional methods for identifying, classifying, or
15 quantifying DNA molecules in a sample of DNA molecules with a
plurality of nucleotide sequences, the method comprising the
steps of digesting said sample with one or more restriction
endonucleases, each said restriction endonuclease recognizing
a subsequence recognition site and digesting DNA to produce
20 fragments with 3' overhangs; contacting said fragments with
shorter and longer oligodeoxynucleotides, each said longer
oligodeoxynucleotide consisting of a first and second
contiguous portion, said first portion being a 3' end
subsequence complementary to the overhang produced by one of
25 said restriction endonucleases, each said shorter
oligodeoxynucleotide complementary to the 3' end of said
second portion of said longer oligodeoxynucleotide stand;
ligating said longer oligodeoxynucleotide to said DNA
fragments to produce a ligated fragment; extending said
30 ligated DNA fragments by synthesis with a DNA polymerase to
form blunt-ended double stranded DNA fragments; amplifying
said double stranded DNA fragments by use of a DNA polymeras
and primer oligodeoxynucleotides to produce amplified DNA
fragments, each said primer oligodeoxynucleotide having a
35 sequence comprising that of a longer oligodeoxynucleotides;
determining the length of th amplified DNA fragments; and
searching a DNA sequence database, said database comprising a

- 40 -

plurality of known DNA sequences that may be present in the
sample, for sequences matching one or more of said fragments
of determined length, a sequence from said database matching
a fragment of determined length when the sequence from said
5 database comprises recognition sites of said one or more
restriction endonucleases spaced apart by the determined
length, whereby DNA sequences in said sample are identified,
classified, or quantified.

      In an eighth embodiment, this invention provides
10 additional methods of detecting one or more differentially
expressed genes in an *in vitro* cell exposed to an exogenous
factor relative to an *in vitro* cell not exposed to said
exogenous factor comprising performing the methods the first
embodiment of this invention wherein said plurality of
15 nucleic acids comprises cDNA of RNA of said *in vitro* cell
exposed to said exogenous factor; performing the methods of
the first embodiment of this invention wherein said plurality
of nucleic acids comprises cDNA of RNA of said *in vitro* cell
not exposed to said exogenous factor; and comparing the
20 identified, classified, or quantified cDNA of said *in vitro*
cell exposed to said exogenous factor with the identified,
classified, or quantified cDNA of said *in vitro* cell not
exposed to said exogenous factor, whereby differentially
expressed genes are identified, classified, or quantified.
25   In a ninth embodiment, this invention provides
additional methods of detecting one or more differentially
expressed genes in a diseased tissue relative to a tissue not
having said disease comprising performing the methods of the
first embodiment of this invention wherein said plurality of
30 nucleic acids comprises cDNA of RNA of said diseased tissue
such that one or more cDNA molecules are identified,
classified, and/or quantified; performing the methods of the
first embodiment of this invention wherein said plurality of
nucleic acids comprises cDNA of RNA of said tissue not having
35 said disease such that one or more cDNA molecules are
identified, classified, and/or quantified; and c mparing said
identified, classified, and/or quantified cDNA molecules of

said diseased tissue with said identified, classified, and/or
quantified cDNA molecules of said tissue not having the
disease, whereby differentially expressed cDNA molecules are
detected.

5   This invention further provides in the ninth
embodiment additional methods wherein the step of comparing
further comprises finding cDNA molecules which are
reproducibly expressed in said diseased tissue or in said
tissue not having the disease and further finding which of

10 said reproducibly expressed cDNA molecules have significant
differences in expression between the tissue having said
disease and the tissue not having said disease, and
optionally wherein said finding cDNA molecules which are
reproducibly expressed and said significant differences in

15 expression of said cDNA molecules in said diseased tissue and
in said tissue not having the disease are determined by a
method comprising applying statistical measures, and
optionally wherein said statistical measures comprise
determining reproducible expression if the standard deviation

20 of the level of quantified expression of a cDNA molecule in
said diseased tissue or said tissue not having the disease is
less than the average level of quantified expression of said
cDNA molecule in said diseased tissue or said tissue not
having the disease, respectively, and wherein a cDNA molecule

25 has significant differences in expression if the sum of the
standard deviation of the level of quantified expression of
said cDNA molecule in said diseased tissue plus the standard
deviation of the level of quantified expression of said cDNA
molecule in said tissue not having the disease is less than

30 the absolute value of the difference of the level of
quantified expression of said cDNA molecule in said diseased
tissue minus the level of quantified expression of said cDNA
molecule in said tissue not having the disease.
          This invention further provides in the ninth

35 embodiment additional methods wherein the diseased tissue and
the tissue not having th disease ar from one or more
mammals, and optionally. wherein the dis ase is a malignancy,

and optionally wherein the disease is a malignancy select d
from the group consisting of prostrate canc r, breast cancer,
colon cancer, lung cancer, skin cancer, lymphoma, and
leukemia.

5   This invention further provides in the ninth
embodiment additional methods wherein the disease is a
malignancy and the tissue not having the disease has a
premalignant character.

           In a tenth embodiment, this invention provides
10 methods of staging or grading a disease in a human individual
comprising performing the methods of the first embodiment of
this invention in which said plurality of nucleic acids
comprises cDNA of RNA prepared from a tissue from said human
individual, said tissue having or suspected of having said
15 disease, whereby one or more said cDNA molecules are
identified, classified, and/or quantified; and comparing said
one or more identified, classified, and/or quantified cDNA
molecules in said tissue to the one or more identified,
classified, and/or quantified cDNA molecules expected at a
20 particular stage or grade of said disease.

           In an eleventh embodiment, this invention provides
additional methods for predicting a human patient's response
to therapy for a disease, comprising performing the methods
of the first embodiment of this invention in which said
25 plurality of nucleic acids comprises cDNA of RNA prepared
from a tissue from said human patient, said tissue having or
suspected of having said disease, whereby one or more cDNA
molecules in said sample are identified, classified, and/or
quantified; and ascertaining if the one or more cDNA
30 molecules thereby identified, classified, and/or quantified
correlates with a poor or a favorable response to one or more
therapies, and optionally which further comprises selecting
one or more therapies for said patient for which said
identified, classified, and/or quantified cDNA molecules
35 correlates with a favorable response.

           In a twelfth embodiment, this invention provides
additional meth ds for evaluating the efficacy of a therapy

- 43 -

in a mammal having a disease, the method comprising
performing the methods of the first embodiment of this
invention wherein said plurality of nucleic acids comprises
cDNA of RNA of said mammal prior to a therapy; performing the
5 method of the first embodiment of this invention wherein said
plurality of nucleic acids comprises cDNA of RNA of said
mammal subsequent to said therapy; comparing one or more
identified, classified, and/or quantified cDNA molecules in
said mammal prior to said therapy with one or more
10 identified, classified, and/or quantified cDNA molecules of
said mammal subsequent to therapy; and determining whether
the response to therapy is favorable or unfavorable according
to whether any differences in the one or more identified,
classified, and/or quantified cDNA molecules after therapy
15 are correlated with regression or progression, respectively,
of the disease, and optionally wherein the mammal is a human.

In a thirteenth embodiment, this invention provides
a kit comprising one or more containers having one or more
restriction endonucleases; one or more containers having one
20 or more shorter oligodeoxynucleotide strands; one or more
containers having one or more longer oligodeoxynucleotide
strands hybridizable with said shorter strands, wherein
either the longer or the shorter oligodeoxynucleotide strands
each comprise a sequence complementary to an overhang
25 produced by at least one of said one or more restriction
endonucleases; and instructions packaged in association with
said one or more containers for use of said restriction
endonucleases, shorter strands, and longer strands for
identifying, classifying, or quantifying one or more DNA
30 molecules in a DNA sample, said instructions comprising (i)
digest said sample with said restriction endonucleases into
fragments, each fragment being terminated on each end by a
recognition site of said one or more restriction
endonucleases; (ii) contact said shorter and longer strands
35 and said digested fragments to form double stranded DNA
adapt rs annealed to said digested fragments, (iii) ligat
said longer strand to said fragments; (iv) gen rate one or

- 44 -

more signals by separating and detecting such of said
fragments that are digest d on each end, each signal
comprising a representation of the length of the fragment and
the identity of the recognition sites on both termini of the

5 fragments; and (v) search a nucleotide sequence database to
determine sequences that match or the absence of any
sequences that match said one or more generated signals, said
database comprising a plurality of known nucleotide sequences
of nucleic acids that may be present in the sample, a

10 sequence from said database matching a generated signal when
the sequence from said database has both (i) the same length
between occurrences of said recognition sites of said one or
more restriction endonucleases as is represented by the
generated signal and (ii) the same recognition sites of said

15 one of more restriction endonucleases as is represented by
the generated signal.

This invention further provides in the thirteenth
embodiment a kit wherein said one or more restriction
endonucleases generate 5' overhangs at the terminus of

20 digested fragments, wherein each said shorter
oligodeoxynucleotide strand consists of a first and second
contiguous portion, said first portion being a 5' end
subsequence complementary to the overhang produced by one of
said restriction endonucleases, and wherein each said longer

25 oligodeoxynucleotide strand comprises a 3' end subsequence
complementary to said second portion of said shorter
oligodeoxynucleotide strand, or optionally wherein said one
or more restriction endonucleases generate 3' overhangs at
the terminus of the digested fragments, wherein each said

30 longer oligodeoxynucleotide strand consists of a first and
second contiguous portion, said first portion being a 3' end
subsequence complementary to the overhang produced by one of
said restriction endonucleases, and wherein each said shorter
oligodeoxynucleotide strand is complementary to the 3' end of

35 said second portion of said longer oligodeoxynucleotide
stand.

This inventi n further provides in the thirteenth
embodiment a kit wherein said instructions further c mprise
those signals expected from one or more DNA molecules of
interest when said sample is digested with a particular one
5 or more restriction endonucleases selected from among said
one or more restriction endonucleases in said kit, and
optionally wherein said one or more DNA molecules of interest
are cDNA molecules differentially expressed in a disease
condition.

10   This invention further provides in the thirteenth
embodiment a kit wherein the restriction endonucleases are
selected from the group consisting of Acc65I, AflII, AgeI,
ApaLI, ApoI, AscI, AvrI, BamHI, BclI, BglII, BsiWI, Bsp120I,
BspEI, BspHI, BsrGI, BssHII, BstYI, EagI, EcoRI, HindIII,
15 MluI, NcoI, NgoMI, NheI, NotI, SpeI, and XbaI.

           This invention further provides in the thirteenth
embodiment a kit further comprising one or more containers
having one or more double stranded adapter DNA molecules
formed by annealing said longer and said shorter
20 oligonucleotide strands.

           This invention further provides in the thirteenth
embodiment a kit further comprising the computer readable
memory of claim 106, or optionally further comprising the
computer readable memory of claim 114, or optionally further
25 comprising the computer readable memory of claim 122.

           This invention further provides in the thirteenth
embodiment a kit further comprising in a container a DNA
ligase, or optionally further comprising in a container a
phosphatase capable of removing terminal phosphates from a
30 DNA sequence.

           This invention further provides in the thirteenth
embodiment a kit further comprising one or more primers, each
said primer consisting of a single stranded
oligodeoxynucleotide comprising the sequence of one of said
35 longer strands; and a DNA polymerase, and optionally wherein
each of said one or more primers further comprises (a) a
first subsequenc that is the portion of the recognition site

- 46 -

of one of said on or more restriction endonucleases
remaining at th terminus of a fragment after digestion, and
(b) a second subsequence of one or two additional nucleotides
contiguous with and 3' to said first subsequence, wherein
5 said primer is detectably labeled such that primers with
differing said one or two additional nucleotides have
different labels that can be distinguishably detected.

This invention further provides in the thirteenth
embodiment a kit wherein said instructions further comprise:
10 detect such of said fragments digested on each end by a
method comprising staining said fragments with silver,
labeling said fragments with a DNA intercalating dye, or
detecting light emission from a fluorochrome label on said
fragments.

15   This invention further provides in the thirteenth
embodiment a kit further comprising reagents for performing a
cDNA sample preparation step; reagents for performing a step
of digestion by one or more restriction endonucleases;
reagents for performing a ligation step; and reagents for
20 performing a PCR amplification step.

4. BRIEF DESCRIPTION OF THE DRAWINGS

These and other features, aspects, and advantages
of the present invention will become better understood by
25 reference to the accompanying drawings, following
description, and appended claims, where:

Fig. 1 illustrates exemplary results of the signals
generated by QEA™ methods of this invention;

Figs. 2A, 2B, and 2C illustrate DNA adapters for an
30 RE/ligation implementation of QEA™ methods of this invention,
where the restriction endonucleases generate 5' overhangs,
open blocks indicating strands of DNA;

Figs. 3A and 3B illustrate the DNA adapters for an
RE/ligation implementation of QEA™ methods of this invention,
35 where the restriction endonucleases generate 3' overhangs;

Figs. 4A, 4B, and 4C illustrate an exemplary biotin
alternative embodiment of QEA™ methods;

- 47 -

Fig. 5 illustrates the DNA prim rs for a PCR embodiment
of QEA™ methods;

Figs. 6A and 6B illustrate a method for DNA sequence
database selection according to this invention;

5   Fig. 7 illustrates an exemplary experimental description
for QEA™ embodiments of this invention;

Figs. 8A and 8B illustrate an overview of a method for
determining a simulated database of experimental results for
QEA™ embodiments of this invention;

10   Fig. 9 illustrates the detail of a method for simulating
a QEA™ reaction;

Figs. 10A-F illustrate exemplary results of the action
of the method of Fig. 9;

Fig. 11 illustrates the detail of a method for
15 determining a simulated database of experimental results for
a QEA™ embodiment of this invention;

Figs. 12A, 12B, and 12C illustrate an exemplary computer
system apparatus, and an alternative embodiment, implementing
methods of this invention;

20   Fig. 13A illustrates exemplary detail of an experimental
design method for QEA™ and CC embodiments of this invention
and Fig. 13B illustrates exemplary detail of an experimental
design method for a QEA™ embodiment of this invention;

Fig. 14 illustrates an exemplary method for ordering the
25 DNA sequences found to be likely causes of a QEA™ signal in
the order of their likely presence in the sample;

Fig. 15 illustrates the detail of a method for
determining a simulated database of experimental results for
a CC embodiment of this invention;

30   Figs. 16A, 16B, 16C, and 16D illustrate exemplary
reaction temperature profiles for preferred manual and
automated implementations of a preferred RE embodiment of a
QEA™ method; and

Figs. 17A-F illustrate the SEQ-QEA™ alternative
35 embodiment of the RE/ligase embodiment of QEA™.

5. DETAILED DESCRIPTI N

According to th pres nt invention, to uniquely
identify an expressed nucleotide or gene sequence, full or
partial, as well as many components of genomic DNA, it is not
5 necessary to determine the actual, complete nucleotide
sequences. Full sequences provide far more information than
is needed to merely classify or determine a sequence
according to this invention. For example, in the human
genome, it is known that there are approximately 10^5 expressed
10 genes. Since the average length of a coding sequence is
approximately 2000 nucleotides, the total number of possible
sequences is approximately 4^{2000}, or about 10^{1200}. The actual
number of expressed human genes is an unimaginably small
fraction (10^{-1195}) of the total number of possible DNA
15 sequences. Even sequencing a 50 bp fragment of a cDNA
sequence generates about 10^{25} times more information than is
needed for classification of that sequence. Use of the
present invention allows direct determination of sequences in
a sample with far less information than either a complete or
20 a partial sequence determination of a sample by making use of
a database of sequences likely to be present in the sample.
If such a database is not available, sequences in the sample
can nevertheless be separately classified.

More generally, the invention is adaptable to
25 analyzing the sequences of any biopolymer, built of a small
number of repeating units, whose naturally occurring
representatives are far fewer that the number of possible,
physical polymers and in which small subsequences can be
recognized. Thus it is applicable to not only naturally
30 occurring DNA polymers but also to naturally occurring RNA
polymers, proteins, glycans, etc.

In computer science, codes which compactly identify
a few members from among a large set of possibilities are
called hash codes. An object of this invention is to
35 construct hash codes for xpressed DNA sequences, or
alternatively for any oth r xisting set of DNA sequences.
In a fully populated hash code without any unassigned code

words, all human genes could be coded by an approximately 17
bit binary number ($2^{17} = 1.3 \times 10^5$). A 20 bit code would be
about 10% filled or 90% sparse ($2^{20} = 1.0 \times 10^6$).

In this invention codes are constructed from one or
5 more signals which represent the presence of short nucleic
acid (preferably DNA) subsequences (hereinafter called
"target subsequences") in the sample sequence and,
preferably, in a QEA™ embodiment, include a representation of
the length along the sample sequence between adjacent target
10 subsequences. In some embodiments, the presence of target
subsequences is directly recognized by direct subsequence
recognition means, including, but not limited to, REs and
other DNA binding proteins, which bind and/or react with
target subsequences, and oligomers of, for example, PNAs or
15 DNAs, which hybridize to target subsequences. In other
embodiments, the presence of effective target subsequences is
recognized indirectly as a result of applying protocols,
perhaps involving multiple DNA binding proteins together with
hybridizing oligomers. In this latter case, each of the
20 multiple proteins or ologomers can recognize a separate
subsequence and the effective target subsequence can be the
combination of the separate subsequences. A preferable
combination is subsequence concatenation in the situation
where all the separately recognized subsequences are
25 adjacent. Such effective target subsequences can have
advantageous properties not achievable by, for example, REs
or PNA oligomers alone. However, this invention, and
particularly its computer methods, are adaptable to any
acceptable subsequence recognition means available in the
30 art. The computer implemented analysis and design methods
treat targer subsequences and effective targer subsequences
in the same manner. Such acceptable subsequence recognition
means preferably precisely and reproducibly recognize target
subsequences and generate a recognition signal with adequate
35 signal to noise ratio and further preferably provide
information on the length betwe n target subsequenc s.

The signals of this invention, which contain
representations of target subsequence occurrences and,
preferably, representations of the length between target
subsequence occurrences, can differ in various embodiments of
5 this invention. In some embodiments, target subsequences are
exactly recognized, for example, where REs are the
recognition means, and subsequence representation can be the
unique identity of the subsequences. In other embodiments,
target subsequence recognition is less exact, for example,
10 where short oligomers are used, and this representation can
be "fuzzy". In the case of short oligomer, a fuzzy
representation can consist of all subsequences which differ
by one nucleotide from a target subsequence, each such
subsequence, perhaps weighted by the probability that each
15 member of the set is the target subsequence. Further, length
representation may depend on the separation and detection
means used to generate the signals. In the case of
electrophoretic separation, the length observed
electrophoretically may need to be corrected, perhaps up to 5
20 to 10%, for mobility differences due to average base
composition differences or due to effects of labeling
moieties used for detection. As these corrections often are
not be known until the total sample sequence is determined,
the length representation of the signal can use the
25 electrophoretic length in bp and not the physical length in
bp. For simplicity and without limitation, in the following
description unless otherwise noted the signals are presumed
to represent physically correct lengths, as if generated by
precise recognition means with a length determined by error
30 or bias free separation and detection means. However, in
particular embodiments, target subsequences can be
represented in a fuzzy manner and length, if present, can
include separation and detection bias.

        Target subsequences recognized are typically
35 contiguous. This is typical for REs adaptable to this
invention. Howev r, this invention is adaptable to means
recognizing discontiguous target subs quences or

discontiguous eff ctive targer subsequences. For example,
oligom rs r cognizing disc ntinuous subsequences can be
constructed by inserting degenerate nucleotides in a
discontinuous region. A set of 16 oligomers recognizing AGC-
5 -TAT, with a two nucleotide discontiguous region, can be
constructed according to the schema TCGNNATA, where N is any
nucleotide. Alternately, such discontiguous subsequences can
be recognized by one oligomer of the form TCGiiATA, where "i"
is inosine, or any other "universal" nucleotide, capable of
10 hybridizing with any naturally occurring base.

          Typically and without limitation, however, the
invention is applied to the analysis of cDNA samples
synthesized from any *in vivo* or *in vitro* sources of RNA.
cDNA can be synthesized either from total cellular RNA, from
15 poly(A)$^+$ RNA, or from specific sub-pools of RNA. Such RNA
sub-pools can be produced by RNA pre-purification, for
example, separation of mRNA of the endoplasmic reticulum from
cytoplasmic mRNA enriches mRNA primarily encoding for cell
surface or extracellular proteins (Celis et al., 1994, *Cell*
20 *Biology*, Academic Press, New York, NY). Such enriched mRNAs
have increased diagnostic or therapeutic utility due, for
example, to their encoded protein's cell-surface or
extracellular roles, such as being a receptor. Such pre-
purified RNA pools can be used in all embodiments of this
25 invention. First strand cDNA synthesis can be performed by
any method known in the art and can use any priming method
known in the art. For example, first strand synthesis
primers can be oligo(dT) primers, random hexamer primers,
phasing primers, mixtures thereof, etc. In particular,
30 phasing primers, containing either an A,C, or G at the 3'
end, can be used in separate cDNA synthesis reactions to
split the cDNA first strands into 3 pools, each generated
from poly(A)$^+$ mRNA having a T, G, or C, respectively, 5' to
the poly(A)$^+$ tail. Twelve pools can be synthesized by using
35 the 12 possible oligo(dT) phasing primers not containing a 3-
terminal thymidin . Further, cDNA can be synthesized by

methods biased to producing full-length cDNAs, e.g. by
requiring pr sence of the 5'-cap in the source cap mRNA.

 Two specific embodiments of the invention are
respectively termed "quantitative expression analysis"
5 ("QEA™") and "colony calling" ("CC"). The specific
embodiment known as QEA™ probes a sample with recognition
means generating signals that preferably comprise an
indication of the presence of a first target subsequence, an
indication of the presence of a second target subsequence,
10 and a representation of the length between the target
subsequences in the sample nucleic acid sequence. If the
first strand of target subsequences occur more than once in a
single nucleic acid in the sample, more than one signal is
generated, each signal comprising the length between adjacent
15 occurrences of the target subsequences.

 QEA™ embodiments are preferred for classifying and
determining sequences in mixtures of cDNAs, but is also
adaptable to samples with only one cDNA. It affords the
relative advantage over prior art methods that cloning of
20 sample nucleic acids is not required. Typically, enough
pairs of target subsequences can be chosen so that sufficient
distinguishable signals can be generated to determine one to
all the sequences in the sample mixture. For example, first,
any pair of target subsequences may occur more than once in a
25 single DNA molecule to be analyzed, thereby generating
several signals with differing lengths from one DNA molecule.
Second, even if a pair of target subsequences occurs only
once in two different DNA molecules to be analyzed, the
lengths between the hits may differ and thus distinguishable
30 signals may be generated.

 The target subsequences used in QEA™ are preferably
optimally chosen by the computer implemented methods of this
invention in view of DNA sequence databases containing
sequences likely to occur in the sample to be analyzed. In
35 the case of human cDNA, efforts of the Human Genome Project
in the United States, efforts abroad, and efforts of private
companies in the sequencing of the human g nome sequences,

both expressed and genetic, are being collected in several
available databases (listed in Sec. 5.1).

Typically, QEA™ can be performed in a "query mode"
or in a "tissue mode." A query mode experiment focuses on
5 determining the expression of a limited number of genes,
perhaps 1 - 100, of interest and of known sequence. A
minimal number of target subsequences are chosen to generate
signals, with the goal that each of the limited number of
genes is discriminated from all the other genes likely to
10 occur in the sample by at least one unique signal. In other
words, such a QEA™ experiment is designed so that each gene
of interest generates at least one signal unique to it (a
"good" gene, see infra). A QEA™ tissue mode experiment
focuses on determining the expression of as many as possible,
15 preferably a majority, of the genes expressed in a tissue or
other sample, without the need for any prior knowledge or
interest in their expression. Target subsequences are
optimally chosen to discriminate the maximum number of sample
DNA sequences into classes comprising one or preferably at
20 most a few sequences. Preferably, enough signals are
produced and detected so that the computer methods of this
invention can uniquely determine the expression of a
majority, or more preferably most, of the genes expressed in
a tissue. In both modes, signals are generated and detected
25 as determined by the threshold and sensitivity of a
particular experiment. Some important determinants of
threshold and sensitivity are the initial amount of mRNA and
thus of cDNA, the amount of molecular amplification performed
during the experiment, and the sensitivity of the detection
30 means.

QEA™ signals are generated by methods comprising a
recognition means for target subsequences that include, but
are not limited to one or more REs in a preferred RE/ligase
embodiment or nucleotide oligomer primers in an alternative
35 PCR embodiment. In both embodiments, this invention
contemplates embodiments which s l ct certain classes of QEA™
reaction products and remove unwanted products. These

embodiments advantageously increase the signal to noise ("s/n") ratio of the resulting signals.

     In general, the RE/ligase method proceeds according to the following steps. The method employs recognition
5 reactions with one, a pair, or more REs which recognize target subsequences with high specificity and cut the sequence at the recognition sites leaving fragments with sticky overhangs characteristic of the particular RE. To each sticky overhang, specially constructed, labeled
10 amplification primers are ligated with the aid of shorter linkers in a manner so that the particular RE making the cut, and thus the particular target subsequence, can be later identified. A DNA polymerase then forms blunt-ended DNA fragments. These fragments are then PCR amplified using the
15 same special labeled primers for a number of cycles preferably just sufficient to detect signals from all fragments of interest and just sufficient to make signals from fragments not of interest, e.g., the linearly amplifying singly cut fragments, relatively insignificant. The
20 amplified labeled fragments are then separated by length using gel electrophoresis in either denaturing or non-denaturing conditions and the length and labeling of the fragments is optically detected. Optionally, single stranded fragments can be removed by a binding hydroxyapatite, or
25 other single strand specific, column or by digestion by a single strand specific nuclease. Also, this invention is adaptable to other functionally equivalent amplification and length separation means. In this manner, the identity of the REs cutting a fragment, and thereby the subsequences present,
30 as well as the length between the cuts is determined.

     The RE/ligase embodiment is adaptable to several embodiments which enhance quantitative characteristics of QEA™ signals or which increase sample sequence discrimination. Certain embodiments use a removal means to
35 improve such quantitative charact ristics as s nsitivity and linear responsiv ness. One or more of th special, labeled amplification primers described abov and used in the PCR

amplification step can have attached removal means comprising
a capture moiety attached to the prim r and a binding partner
attached to a solid support, e.g., biotin and streptavidin
beads. In this manner certain products of the PCR reactions,
5 e.g., fragments cut with different REs at each end, can be
separated and purified from background fragments. Such
purified fragments can thereby be detected with increased
sensitivity. For example, fragments cut with pairs of
different REs on both ends are preferably separated since
10 such fragments contain the majority of signals. With N REs,
there are (N-1)/2 pairs with different REs but only N pairs
with the same RE.

Alternatively, cDNA is synthesized from an mRNA
sample with synthesis primers at least one of which is
15 biotinylated. In the case where only one synthesis primer is
biotinylated, the cDNA is then cyclized. In any case, the
cDNA is then cut with a one or a pair of REs, and the
special, labeled amplification primers are ligated to the cut
ends with the aid of shorter linkers as previously discussed.
20 The singly cut ends attached to the biotinylated cDNA
synthesis primers are removed with streptavidin or avidin
beads leaving highly pure double cut cDNA fragments with
ligated amplification primers, but with minimal singly cut
and labeled background fragments. With sufficiently
25 sensitive detection means, these pure doubly cut and labeled
fragments can be directly detected, after separation by
length (e.g., by electrophoresis or column chromatography),
without amplification. If amplification is needed, absence
of the DNA singly cut background fragments improves signal t
30 ncise ratio resulting in fewer necessary amplification
cycles. Thereby, PCR amplification bias is decreased or
eliminated and linear responsiveness of QEA™ signals to input
mRNA amounts is improved.

Other RE/ligase embodiments increase sample
35 sequence discrimination in QEA™ experiments, for example, by
recognizing target subsequences longer or less limited than
those recogniz d by REs, or by recognizing third subsequ nc s

interior to cut fragments. This added information can often
discriminate two sample sequences producing fragments having
identical original end subsequences and lengths. It is used
in the computer implemented database lookup methods of this
5 invention in a manner similar to the use of target
subsequences. In one embodiments, the target subsequences
recognized can be effectively lengthened by using an
amplification primer with an internal Type IIS RE recognition
site so positioned that the Type IIS RE cuts the amplified
10 fragments in a manner producing a second overhang contiguous
with the recognition site of the initial RE. The sequence of
the second overhang concatenated with the initial target end
subsequence produces an effectively longer target
subsequence. Alternatively, an effectively longer target
15 subsequence can be recognized by using phasing primers during
PCR amplification. The PCR amplification step can de divided
into several pools with each pool using one phasing
amplification primer constructed so as to recognize one or
more additional nucleotides beyond the original RE
20 recognition site. These additional nucleotides then
contribute to an effectively longer target subsequence.

        A third subsequence internal to a fragment can be
recognized by a distinctively labeled probe binding or
hybridizing with the third subsequence. Such a probe added
25 before detection generates unique signals from the fragment
containing that subsequence. Alternatively, a probe can
suppress signals from fragments with the third subsequence.
For example, a probe added before the PCR amplification step
and which prevents amplification of a fragment with the third
30 subsequence thereby removes and suppresses any signal from
such fragments. Such a probe can be without limitation
either an RE for recognizing and cutting the fragment with
the third subsequence or a PNA or modified DNA oligomer,
which cannot serve as a PCR primer, for hybridizing with the
35 third subsequence. Also, a third subsequence can be the
sequence f th overhang produced by a Type IIS RE cutting
the amplification primers sufficiently close to their 3' nds

so that the resulting overhang is not contiguous with the
recogniti n sequ nce of the initial RE.

Further, various embodiments for improving the
quantitative characteristics of QEA™ experiments and for
5 improving the discrimination of sample sequences can be
combined in advantageous fashions to achieve both
improvements in the same experiment. For example, removal
means to increase the s/n ratio is combined with a Type IIS
RE cutting the amplification primers to increase sample
10 sequence discrimination in an embodiment called SEQ-QEA™.

In a preferred PCR method for QEA™, a suitable
collection of target subsequences is chosen by the computer
implemented QEA™ experimental design methods, and PCR primers
distinctively labeled with fluorochromes are synthesized to
15 hybridize with these target subsequences. The primers are
designed as described in Sec. 5.3 to reliably recognize short
subsequences while achieving a high specificity in PCR
amplification. Using these primers, a minimum number of PCR
amplification steps amplifies those fragments between the
20 primed subsequences existing in DNA sequences in the sample,
thereby recognizing the target subsequences. The labeled,
amplified fragments are then separated by gel electrophoresis
and detected. Further, the PCR embodiment is adaptable to
the same embodiment previously discussed with respect to the
25 RE/ligase embodiment.

The signals generated from the recognition
reactions of a QEA™ experiment are analyzed by computer
methods of this invention. The analysis methods simulate a
QEA™ experiment using a database either of substantially all
30 known DNA sequences or of substantially all, or at least a
majority of, the DNA sequences likely to be present in a
sample to be analyzed and a description of the reactions to
be performed. The simulation results in a digest database
which contains for each possible signal that can be generated
35 the database sequences responsible for that signal. Thereby,
finding the sequences that can g nerate a signal involves a
look-up in the simulated digest database. Computer

- 58 -

implemented design methods optimize the choice of target
subsequences in QEA™ r actions in order to maximize the
information produced in an experiment. For the tissue mode,
the methods maximize the number of sequences having unique
5 signals by which their quantitative presence can be
unambiguously determined. For the query mode, the methods
maximize only the number of sequences of interest having
unique signals, ignoring recognition of other sequences that
might be present in a sample.
10   The second specific embodiment known as colony
calling ("CC") generates subsequence occurrence data without
length information. Since this method requires only
hybridizations, it is preferred for gene identification in
arrayed single-sequence clones constructed from a tissue
15 library. This embodiment constructs a binary code in which
each bit of the code represents the presence or absence of
one target subsequence. By probing four to eight target
subsequences in parallel, such as by using distinguishable
fluorescent labeling of the multiple probes, in view of the
20 adequacy of a 20 bit code, the presence or absence of any
expressed human gene should be determinable in just three to
five separate probe steps. Such a compact method with such
economy in signal generation is highly useful.
Alternatively, recent real time hybridization detection
25 methods (Stimson et al., 1995, Proc. Natl. Acad. Sci. USA,
92:6379-6383) based on optical wave guides can be used for
detection. These methods make hybridization detection more
efficient both by eliminating the washing step otherwise
needed between hybridization and detection and by speeding up
30 the detection step.
           The hash code generated by the probe hybridization
reactions is interpreted by computer implemented methods of
this invention. The analysis methods simulate a CC
experiment using a list of the target subsequences and a
35 database of the DNA sequences likely to be present in a
sample to be analyzed. The simulation results in a hash code
table which contains for each hash code all possible

s quences that can generat that code. Thereby,
interpretation of a det cted hash code requires a look-up in
the table to find the possible sequences.

It is preferable that subsequences be carefully
5 chosen in order that a minimum set of targets be obtained,
preferably no more than approximately 20, that produce the
maximum amount of information. Computer implemented methods
of this invention determine optimum sets of target
subsequences for a given database of sequences likely to
10 occur in the sample by optimizing the number of non-empty
hash codes in the simulated hash code table.

Maximum information is obtained when the target
subsequences occur completely randomly in the possible sample
sequences, that is, when their likelihood of occurrence is
15 approximately 50% and the presence of one subsequence is
independent of the presence of any other subsequence.
Therefore, target subsequences chosen to generate a signal
should preferably occur in the DNA sequence sample to be
analyzed less than about 50% and at least more often than 5-
20 10%, preferably more often than 10-15%. The most preferable
occurrence probability is from 25-50%. Also the presence of
one target subsequence is preferably probabilistically
independent of the presence of any other subsequence.

Using data on expressed RNA from human DNA sequence
25 databases, this means that sub-sequences are preferably less
than about 5 to 8 bp long for cDNA classification.
Typically, the resulting preferable target subsequences are 4
to 6 bp long. Longer sequences occur too infrequently to be
preferred for use. However, for classifying gDNA, longer
30 subsequences, up to 20 to 40 bp, are preferably used, because
gDNA fragments are normally of much greater length, from at
least 5 kilobases ("kb") for plasmid inserts to more the 100
kb for P1 inserts, and thus would typically have more
sequence variability, requiring longer target subsequences.
35   The preferred hybridization probes for short target
subsequenc s are labeled peptid -nucl ic acids (PNAs).
Alternatively sets of degenerate, longer DNA oligonucleotides

- 60 -

are used which include as a common subsequence the target
subsequence. These degenerate sets achieve improved
hybridization specificity as compared to 4 to 6-mers. Sets
of probes, each probe distinctively and distinguishably
5 labeled with a fluorochrome, are hybridized in conditions of
high stringency to arrayed DNA sequence clones and optically
detected to detect the presence of target subsequences. For
example, in an embodiment wherein five fluorochromes are
simultaneously distinguished and 20 subsequences observations
10 are required for gene identification (a 20 bit code), any
gene in a colony can be identified in only four hybridization
steps. Alternately, efficient hybridization detection means
based on optical wave guide detection of DNA hybridization
can be used. By using differently sized and .shaped particles
15 associated with different probes, the resultant differences
in light scattering can be used to detect hybridization of
multiple probes simultaneously with these wave guide methods.

        Target subsequences can be chosen tc discriminate
not only single genes but also, more coarsely, sets of genes.
20 Fewer target subsequences can be chosen so that a particular
pattern of hits will indicate the presence of a gene of a
particular type. Types of genes of interest might be
oncogenes, tumor suppressor genes, growth factors, cell cycle
genes, or cytoskeletal genes, etc.
25   In embodiments of this invention where high
stringency hybridization are specified, such conditions
generally comprise a low salt concentration, equivalent to a
concentration of SSC (173.5 g. NaCl, 88.2 g. Na Citrate, H_2O
to 1 l.) of less than approximately 1 mM, and a temperature
30 near or above the T_m of the hybridizing DNA. In contrast,
conditions of low stringency generally comprise a high salt
concentration, equivalent to a concentration of SSC of
greater than approximately 150 mM, and a temperature below
the T_m of the hybridizing DNA.
35   In embodiments of this invention where DNA
oligomers are specified for performing functions, including
hybridization and chain el ngation priming, alternatively

- 61 -

oligomers can be used that comprise thos of the following
nucleotid mimics which perform similar functions.
Nucleotide mimics are subunits (other than classical
nucleotides) which can be polymerized to form molecules
5 capable of specific, Watson-Crick-like base pairing with DNA.
The oligomers can be DNA or RNA or chimeric mixtures or
derivatives or modified versions thereof. The oligomers can
be modified at the base moiety, sugar moiety, or phosphate
backbone. The oligomers may include other appending groups
10 such as peptides, hybridization-triggered cleavage agents
(see, e.g., Krol et al., 1988, *BioTechniques* 6:958-976), or
intercalating agents (see, e.g., Zon, 1988, *Pharm. Res.* ,
5:539-549). The oligomers may be conjugated to another
molecule, e.g., a peptide, hybridization triggered cross-
15 linking agent, transport agent, hybridization-triggered
cleavage agent, etc.

The oligomers may also comprise at least one
nucleotide mimic that is a modified base moiety which is
selected from the group including, but not limited to,
20 5-fluorouracil, 5-bromouracil, 5-chlorouracil, 5-iodouracil,
hypoxanthine, xantine, 4-acetylcytosine,
5-(carboxyhydroxylmethyl) uracil, 5-carboxymethylaminomethyl-
2-thiouridine, 5-carboxymethylaminomethyluracil,
dihydrouracil, beta-D-galactosylqueosine, inosine,
25 N6-isopentenyladenine, 1-methylguanine, 1-methylinosine,
2,2-dimethylguanine, 2-methyladenine, 2-methylguanine,
3-methylcytosine, 5-methylcytosine, N6-adenine,
7-methylguanine, 5-methylaminomethyluracil,
5-methoxyaminomethyl-2-thiouracil, beta-D-mannosylqueosine,
30 5'-methoxycarboxymethyluracil, 5-methoxyuracil,
2-methylthio-N6-isopentenyladenine, uracil-5-oxyacetic acid
(v), wybutoxosine, pseudouracil, queosine, 2-thiocytosine,
5-methyl-2-thiouracil, 2-thiouracil, 4-thiouracil,
5-methyluracil, uracil-5-oxyacetic acid methylester,
35 3-(3-amino-3-N-2-carboxypropyl) uracil, and
2,6-diaminopurin . The olig mers may comprise at least one
modified sugar moiety selected from the group including but

- 62 -

not limited to arabinose, 2-fluoroarabinose, xylulose, and
hexose. The oligom rs may comprise at least one modified
phosphate backbone selected from the group consisting of a
phosphorothioate, a phosphorodithioate, a
5 phosphoramidothioate, a phosphoramidate, a phosphordiamidate,
a methylphosphonate, an alkyl phosphotriester, and a
formacetal or analog thereof.

        The oligomer may be an α-anomeric oligomer. An α-
anomeric oligomer forms specific double-stranded hybrids with
10 complementary RNA in which, contrary to the usual β-units,
the strands run parallel to each other (Gautier et al., 1987,
Nucl. Acids Res. 15:6625-6641).

        Oligomers of the invention may be synthesized by
standard methods known in the art, *e.g.*, by use of an
15 automated DNA synthesizer (such as are commercially available
from Biosearch, Applied Biosystems, etc.). As examples,
phosphorothioate oligos may be synthesized by the method of
Stein et al. (1988, *Nucl. Acids Res.* 16:3209),
methylphosphonate oligos can be prepared by use of controlled
20 pore glass polymer supports (Sarin et al., 1988, *Proc. Natl.
Acad. Sci. U.S.A.* 85:7448-7451), etc.

        In specific embodiments of this invention it is
preferable to use oligomers that can specifically hybridize
to subsequences of a DNA sequence too short to achieve
25 reliably specific recognition, such that a set of target
subsequences is recognized. Further where PCR is used, as
Taq polymerase tolerates hybridization mismatches, PCR
specificity is generally less than hybridization specificity.
Where such oligomers recognizing short subsequences are
30 preferable, they may be constructed in manners including but
not limited to the following. To achieve reliable
hybridization to shorter DNA subsequences, degenerate sets of
DNA oligomers may be used which are constructed of a total
length sufficient to achieve specific hybridization with each
35 member of the set containing a shorter sequence complementary
to the common subsequence to be recognized. Alternatively, a
longer DNA oligomer may be constructed with a shorter

sequence c mplementary to the subsequence to be recognized
and with additional universal nucleotides or nucleotide
mimics, which are capable of hybridizing to any naturally
occurring nucleotide. Nucleotide mimics are sub-units which
5 can be polymerized to form molecules capable of specific,
Watson-Crick-like base pairing with DNA. Alternatively, the
oligomers may be constructed from DNA mimics which have
improved hybridization energetics compared to naturally
occurring nucleotides.
10   A preferred mimic is a peptido-nucleic acid ("PNA")
based on a linked N-(2-aminoethyl)glycine backbone to which
normal DNA bases have been attached (Egholm et al., 1993,
Nature 365:566-67). This PNA obeys specific Watson-Crick
base pairing, but with greater free energy of binding and
15 correspondingly higher melting temperatures. Suitable
oligomers may be constructed entirely from PNAs or from mixed
PNA and DNA oligomers.
          In embodiments of this invention where DNA
fragments are separated by length, any length separation
20 means known in the art can be used. One alternative
separation means employs a sieving medium for separation by
fragment length coupled with a force for propelling the DNA
fragments though the sieving medium. The sieving medium can
be a polymer or gel, such a polyacrylamide or agarose in
25 suitable concentrations to separate 10-1000 bp DNA fragments.
In this case the propelling force is a voltage applied across
the medium. The gel can be disposed in electrophoretic
configurations comprising thick or thin plates or
capillaries. The gel can be non-denaturing or denaturing.
30 Alternately, the sieving medium can be such as used for
chromatographic separation, in which case a pressure is the
propelling force. Standard or high performance liquid
chromatographic ("HPLC") length separation means may be used.
An alternative separation means employs molecular
35 characteristics such as charge, mass, or charge to mass
ratio. Mass spectrographic means capable of separating 10-
1000 bp fragm nts may be us d.

DNA fragment lengths determined by such a
separation means represent the physical length in base pairs
between target subsequences, after adjustment for biases or
errors introduced by the separation means and length changes
5 due to experimental variables (e.g., presence of a detectable
label, ligation to an adapter molecule). A represented
length is the same as the physical length between occurrences
of target subsequences in a sequence from said database when
both said lengths are equal after applying corrections for
10 biases and errors in said separation means and corrections
based on experimental variables. For example, represented
lengths determined by electrophoresis can be adjusted for
mobility biases due to average base composition or mobility
changes due to an attached labeling moiety and/or adapter
15 strand by conventional software programs, such as Gene Scan
Software from Applied Biosystems, Inc. (Foster City, CA).

In embodiments of this invention where DNA
fragments must be labeled and detected, any compatible
labeling and detection means known in the art can be used.
20 Advances in fluorochromes, in optics, and in optical sensing
now permit multiply labeled DNA fragments to be distinguished
even if they completely overlap in space, as in a spot on a
filter or a band in a gel. Results of several recognition
reactions or hybridizations can be multiplexed in the same
25 gel lane or filter spot. Fluorochromes are available for DNA
labeling which permit distinguishing 6-8 separate products
simultaneously (Ju et al., 1995, *Proc. Natl. Acad Sci. USA*
<u>92</u>:4347-4351).

Exemplary fluorochromes adaptable to this invention
30 and methods of using such fluorochromes to label DNA are
described in Sec. 6.11.

Single molecule detection by fluorescence is now
becoming possible (Eigen et al., 1994, *Proc. Natl. Acad Sci.
USA* <u>91</u>:5740-5747), and can be adapted for use.
35   In embodiments of this invention where
intercalating DNA dy s are utilized to detect DNA, any such
dye known in the art is adaptable. In particular such dyes

include but are not limited to ethidium bromide, propidium
iodide, Hoechst 33258, Hoechst 33342, acridine orange, and
ethidium bromide homodimers. Such dyes also include POPO,
BOBO, YOYO, and TOTO from Molecular Probes (Eugene, OR).

5   Finally alternative sensitive detection means
available include silver staining of polyacrylamide gels
(Bassam et al., 1991, *Analytic Biochemistry* 196:80-83), and
the use of intercalating dyes. In this case the gel can be
photographed and the photograph scanned by scanner devices
10 conventional in the computer art to produce a computer record
of the separated and detected fragments. A further
alternative is to blot an electrophoretic separating gel onto
a filter (e.g., nitrocellulose) and then to apply any
visualization means known in the art to visualize adherent
15 DNA. See, e.g., Kricka et al., 1995, Molecular Probing,
Blotting, and Sequencing, Academic Press, New York. In
particular, visualization means requiring secondary reactions
with one or more reagents or enzymes can be used, as can any
means employed in the CC embodiment.

20   A preferred separation and detection apparatus for
use in this invention is found in copending U.S. Patent
Application Serial No. 08/438,231 filed May 9, 1995, which is
hereby incorporated by reference in its entirety. Other
detection means adaptable to this invention include the
25 commercial electrophoresis machines from Applied Biosystems
Inc. (Foster City, CA), Pharmacia (ALF), Hitachi, Licor. The
Applied Biosystems machine is preferred among these as it is
the only machine capable of simultaneous 4 dye resolution.

               In the following subsections and the accompanying
30 examples sections QEA™ and the CC embodiments are described
in detail.

5.1. QUANTITATIVE EXPRESSION ANALYSIS

This embodiment of this invention in the tissue
35 mode preferably generates one or more signals unique to each
cDNA sequence in a mixture of cDNAs, such as may be deriv d
fr m total cellular RNA or total cellular mRNA from a tissue

sample, and to quantitativ ly relate the str ngth of such a
signal or signals to the relative amount of that cDNA
sequence in the sample or library. In the query mode, this
embodiment preferably generates signals uniquely
5 discriminating only a few sample sequences of interest in a
quantitative manner. Less preferably, the signals uniquely
determine only sets of a small number of sequences, typically
2 - 10 sequences. QEA™ signals comprise an indication of the
presence of pairs of target subsequences and the length
10 between pairs of adjacent subsequences in a DNA sample.
Alternatives include recognizing the presence of third
subsequences between the pairs of target subsequences. In a
further embodiment ("5'-QEA™"), one of the subsequences is
the true end of the protein coding sequence, in a defined
15 relation to the 5' cap of the source mRNA. Signals are
preferably generated in a manner permitting straightforward
automation with existing laboratory robots. For simplicity
of disclosure, and not by way of limitation, the detailed
description of this method is directed to the analysis of
20 samples comprising a plurality of cDNA sequences. It is
equally applicable to samples comprising a single sequence or
samples comprising sequences of other types of DNA or nucleic
acids generally.

While described in terms of cDNA hereinbelow, it
25 will be understood that the DNA sample can be cDNA and/or
genomic DNA, and preferably comprises a mixture of DNA
sequences. In specific embodiments, the DNA sample is an
aliquot of cDNA of total cellular RNA or total cellular mRNA,
most preferably derived from human tissue. The human tissue
30 can be diseased or normal. In one embodiment, the human
tissue is malignant tissue, e.g., from prostate cancer,
breast cancer, colon cancer, lung cancer, lymphatic or
hematopoietic cancers, etc. In another embodiment, the
tissue may be derived from in vivo animal models of disease
35 or other biologic processes. In this cases the diseases
m deled can usefully includ , as well as cancers, diabetes,
obesity, the rheumatoid or autoimmune diseases, etc. In yet

another   mbodiment, the samples can be derived from *in vitro*
cultures and models. This invention can also be
advantageously applied to examine gene expression in plants,
yeasts, fungi, etc.

5   The cDNA, or the mRNA from which it is synthesized,
must be present at some threshold level in order to generate
signals, this level being determined to some degree by the
conditions of a particular QEA™ experiment. For example,
such a threshold is that preferably at least 1000, and more
10 preferably at least 10,000, mRNA molecules of the sequence to
be detected be present in a sample. In the case where one or
only a few mRNAs of a type of interest are present in each
cell of a tissue from which it is desired to derive the
sample mRNA, at least a corresponding number of such cells
15 should be present in the initial tissue sample. In a
specific embodiment, the mRNA detected is present in a ratio
to total sample RNA of $1:10^5$ to $1:10^6$. With a lower ratio,
more molecular amplification can be performed during a QEA'"
experiment.

20   The cDNA sequences occurring in a tissue derived
pool include short untranslated sequences and translated
protein coding sequences, which, in turn, may be a complete
protein coding sequence or some initial portion of a coding
sequence, such as an expressed sequence tag. A coding
25 sequence may represent an as yet unknown sequence or gene or
an already known sequence or gene entered into a DNA sequence
database. Exemplary sequence databases include those made
available by the National Center for Biotechnology
Information ("NCBI") (Bethesda, MD) (GenBank) and by the
30 European Bioinformatics Institute ("EMBL") (Hinxton Hall,
UK).

A QEA™ method is also applicable to samples of
genomic DNA in a manner similar to its application to cDNA.
In gDNA samples, information of interest includes occurrence
35 and identity of translocations, gene amplifications, loss of
heterozygosity for an allel , etc. This information is of
interest in cancer diagnosis and staging. In cancer

patients, amplified sequences might reflect an oncogene,
whil loss of heterozygosity might reflect a tumor suppress r
gene. Such sequences of interest can be used to select
target subsequences and to predict signals generated by a
5 QEA™ experiment. Even without prior knowledge of the
sequences of interest, detection and classification of QEA™
signal patterns is useful for the comparison of normal and
diseased states or for observing the progression of a disease
state. Gene expression information concerning the
10 progression of a disease state is useful in order to
elucidate the genetic mechanisms behind disease, to find
useful diagnostic markers, to guide the selection and observe
the results of therapies, etc. Signal differences identify
the gene or genes involved, whether already known or yet to
15 be sequenced.

        Classification of QEA™ signal patterns, in an
exemplary embodiment, can involve statistical analysis to
determine significant differences between patterns of
interest. This can involve first grouping samples that are
20 similar in one of more characteristics, such characteristics
including, for example, epidemiological history,
histopathological state, treatment history, etc. Signal
patterns from similar samples are then compared, e.g., by
finding the average and standard deviation of each individual
25 signals. Individual signals which are of limited
variability, for which the standard deviation is less than
the average, then represent genetic constants of samples of
this particular characteristic. Such limited variability
signals from one set of tissue samples can then be compared
30 to limited variability signals from another set of tissue
samples. Signals which differ in this comparison then
represent significant differences in the genetic expression
between the tissue samples and are of interest in reflecting
the biological differences between the samples, such as the
35 differences caused by the progression of a disease. For
example, a significant difference in expressi n is detect d
with the difference in the genetic expression b tween two

tissues exceed the sum of the standard deviation of the
expressions in the tissues. Other standard statistical
comparisons can also be used to establish level of expression
and the significance of differences in levels of expressions.

5   Target subsequence choice is important in the
practice of this invention. The two primary considerations
for selecting subsequences are, first, redundancy, that is,
that there be enough target subsequence pair occurrences
(also known as "hits") per gene that a unique signal is

10 likely to be generated for each sample sequence, and second,
resolution, that is, that there not be so target subsequence
pair occurrences with very similar lengths in a sample that
the signals cannot be resolved. For sufficient redundancy,
it is preferable that there be on average, approximately

15 three target subsequence pair hits per gene or DNA sequence
in the sample. It is highly preferable that there be a
minimum of at least one pair hit per each gene In tests of a
database of eukaryotic expressed sequences, it has been found
that an average value of three pair hits per gene appears to

20 be generally a sufficient guarantee of this minimum
criterion.

          Sufficient resolution depends on the separation and
detection means chosen. For a particular choice of
separation and detection means, a recognition reaction

25 preferably should not generate more fragments than can be
separated and distinguishably detected. In a preferred
embodiment, gel electrophoresis is the separation means used
to separate DNA fragments by length. Existing
electrophoretic techniques allow an effective resolution of

30 three base pair ("bp") length differences in sequences of up
to 1000 bp length. Given knowledge of fragment base
composition, effective resolution down to 1 bp is possible by
predicting and correcting for the small differences in
mobility due to differing base composition. However and

35 without limitation, an easily achievable three bp resolution
is assumed by way of example in the description of the
invention herein. It is preferable for incr ased detecti n

efficiency that the distinguishably labeled products from as
many recognition r actions as possible be combined for
separation in one gel lane. This combination is limited by
the number of labels distinguishable by the employed
5 detection means. Any alternative means for separation and
detection of DNA fragments by length, preferably with
resolution of three bp or better, can be employed. For
example, such separation means can be thick or thin plate or
column electrophoresis, column chromatography or HPLC, or
10 physical means such as mass spectroscopy.

The redundancy and resolution criteria are
probabilistically expressed in Eqns. 1 and 2 in an
approximation adequate to guide subsequence choice. In these
equations the number of genes in the cDNA sequence mixture is
15 N, the average gene length is L, the number of target
subsequence pairs is M (the number of pairs of recognition
means), and the probability of each target subsequence
occurring in, or hitting, a typical sample sequence is p.
Since each target subsequences is preferably selected to
20 occur independently in each sample sequence, the probability
of occurrence of an arbitrary subsequence pair is then p^2.
Eqn. 1 expresses the redundancy condition of three pair
occurrences per sample sequence, assuming the probability of
occurrence of each target subsequence is independent.
25

$$Mp^2 = 3 \tag{1}$$

Eqn 2 expresses the resolution condition of having fragments
with lengths no closer on average than 3 base pairs. This
equation approximates the actual fragment length distribution
30 with a uniform distribution.

$$\frac{L}{Np^2} = 3 \tag{2}$$

Given expected values of N, the number of sequences in the
35 library or sample to analyze (library complexity), and L, the
average expr ssed sequ nce (or gene) length, Eqns 1 and 2 are

solved for the subsequence occurrence probability and number
of subsequ nces required. This solution d pends on the
particular redundancy and resolution criteria dictated by the
particular experimental method chosen to implement QEA™.
5 Alternative values may be required for other implementations
of this embodiment.

For example, it is estimated that the entire human
genome contains approximately 10^5 protein coding sequences
with an average length of 2000. The solution of Eqns 1 and 2
10 for these parameters is p = 0.082 and M = 450. Thereby the
expression of all genes in all human tissues can be analyzed
with 450 target subsequence pairs, each subsequence having an
independent probability of occurrence of 8.2%. In an
embodiment in which eight fluorescently labeled subsequence
15 pairs can be optically distinguished and detected per
electrophoresis lane, such as is possible when using the
separation and detection apparatus described in copending
U.S. Patent Application Serial No. 08/438,231 filed May 9,
1995, 450 reactions can be analyzed in only 57 lanes.
20 Thereby only one electrophoresis plate is needed in order to
completely determine all human genome expression levels.
Since the best commercial machines known to the applicants
can discriminate only four fluorescent labels in one lane, a
corresponding increase in the number of lanes is required to
25 perform a complete genome analysis with such machines.

As a further example, it is estimated that a typically
complex human tissue expresses approximately 15,000 genes.
The solution for N = 15000 and L = 2000 is p = 0.21 and M =
68. Thus expression in a typical tissue can be analyzed with
30 68 target subsequence pairs, each subsequence having an
independent probability of occurrence of 21%. Assuming 4
subsequence pairs can be run per gel electrophoresis lane,
the 68 reactions can be analyzed in 17 lanes in order to
determine the gene expression frequencies in any human
35 tissue. Thus it is clear that this method leads to greatly
simplified quantitative g n xpressi n analysis within the
capabilities of existing lectrophoretic systems.

These equations provide an adequate guide to picking subsequence pairs. Typically, preferred probabilities of target subsequence occurrence are from approximately 0.01 to 0.30. Probabilities of occurrence of
5 specified subsequences and RE recognition sites can be determined from databases of DNA sample sequences. Example 6.2 lists these probabilities for exemplary RE recognition sites. Appropriate target subsequences can be selected from these tables. Computer implemented QEA™ experimental design
10 methods can then optimize this initial selection.

Another use of QEA™ is to compare directly the expression of only a few genes or sample sequences, typically 1 to 10, between two different tissues, the query mode, instead of seeking to determine the expression of all genes
15 in a tissue, the tissue mode. In this query mode, a few target subsequences are selected to discriminate the genes of interest both among themselves and from all other sequences possibly present. The computer design methods described hereinbelow can make this selection. If 4 subsequence pairs
20 are sufficient for identification, then the fragments from the 4 recognition reactions performed on each tissue are preferably separated and detected on two separate lanes in the same gel. If 2 subsequence pairs are sufficient for identification, the two tissues are preferably analyzed in
25 the same gel lane. Such comparison of signals from the same gel improves quantitative results by eliminating measurement variability due to differences between separate electrophoretic runs. For example, expression of a few target genes in diseased and normal tissue samples can be
30 rapidly and reliably analyzed.

The query mode of QEA™ is also useful even if the sequences of the particular genes of interest are not yet known. Differentially expressed features can be identified by comparing the results of QEA™ reactions applied to two
35 different samples. In the case where the separation and detection of r action products is by gel electrophoresis, such a comparison can be done by comparing gel bands or

fluorescent traces of exiting fragments. Such differentially express d features can then retrieved from the gel by methods known in the art (e.g., electro-elution from the gel) and the DNA fragments analyzed by conventional techniques, such as by

5 sequencing. Such sequences, which are typically partial, can then be used as probes (e.g., in PCR or Southern blot hybridization) to recover full-length sequences. In this manner, QEA™ techniques can guide the discovery of new differentially expressed cDNA or of changes of the state of

10 gDNA. The sequences of the newly identified genes, once determined, can then be used to guide QEA™ target subsequence choice for further analysis of the differential expression of the new genes.

Alternative embodiments of QEA™ are described

15 herein, differing primarily in how the recognition means recognize the target subsequences. Associated with these primary differences are secondary differences in how signals are generated from the recognition means. In the PCR embodiment, target subsequences are recognized by oligomers

20 which hybridize to the DNA target subsequences and act as PCR primers for the amplification of the segments between adjacent primer pairs. Amplified fragments from a sample are separated preferably by electrophoresis. Selection of target subsequences, the primer hybridizing sites, meeting the

25 probability of occurrence and independence criteria is preferably made from a database containing sequences expected to be present in the samples to be analyzed, for example human GenBank sequences, and optimized by the computer implemented experimental design methods. In a preferred

30 embodiment, subsequence selection begins by compiling oligomer frequency tables containing the frequencies of, preferably, all 4 to 8-mers by using a sequence database. From these tables, target subsequences with the necessary probabilities of occurrence according to Eqns. 1 and 2 are

35 selected and checked for independence, by, for example, checking that the conditional pr bability for occurrence by any selected pair of subsequences is the product f th

probabilities of occurrences of the individual subsequences
f the pair. An initial selecti n can be optimized to
determine target subsequence sets producing unique fragments
from the greatest number of sample sequences. PCR primers
5 are synthesized with a 3' end complementary to the chosen
subsequences and used in the PCR embodiment. Example 6.1
illustrates the signals output by this method in a specific
example.

        The preferred embodiment uses DNA binding proteins,
10 specifically REs, including Type IIS REs, to recognize and
cleave sample sequences at the target subsequences. Desired
fragments, with lengths dependent only on source cDNA
sequence, are amplified by an amplification means in order to
dilute remaining, unwanted fragments with indefinite lengths.
15 Typically, but without limitations, desired fragments are
doubly cut by REs whereas unwanted fragments are singly cut.
But in 5'-QEA™, singly cut fragments have a definite length
and are of interest. Unwanted singly cut fragments can be
removed by affinity means (e.g., biotin labeling), physical
20 means (e.g., hydroxyapatite column separation), or enzymatic
means (e.g., single strand specific nucleases). Sufficient
removal of the unwanted singly cut ends from the desired
doubly cut fragments can permit fragment detection without an
amplification step. For the RE alternative embodiments, the
25 possible target subsequences, although limited to recognition
sites of available REs, can be selected in a manner similar
to the above in order to meet the previous probability or
occurrence and independence criteria as closely as possible.
For example, the probabilities of occurrence of various RE
30 recognition sites can be determined from a database of
potential sample sequences, and those REs chosen with
recognition subsequences whose probabilities of occurrence
meet the criterion of Eqns 1 and 2 as closely as possible.
If multiple REs satisfy the selection criteria, a subset is
35 selected by including only those REs with independently
occurring recognition subsequ nces, determin d, for example,
in the previ us manner using conditional probabilities of

- 75 -

occurrence. An initial choice can be optionally optimized by
th computer impl mented experimental d sign m thods.

A number, R_e, of REs are preferably s lected so that
the number of RE pairs is approximately M, as determined from
5 Eqn. 1, where the relation between M and R_e is given by Eqn.
3.

$$M = \frac{R_e (R_e + 1)}{2} \tag{3}$$

10 For example, a set a set of 20 acceptable REs results in 210
subsequence pairs.

There are numerous REs currently available, whose
recognition sequences have a wide range of occurrence
probabilities, from which REs can be selected for the present
15 invention. Exemplary REs are listed in Sec. 6.2.

The PCR and the RE embodiments have different
accuracy and flexibility characteristics. RE embodiments are
generally more accurate, with fewer false positive and false
negative identifications, since the enzymatic recognition and
20 subsequent ligation reactions are generally more specific
than the hybridization of short PCR primers to their
subsequence targets, even under stringent hybridization
conditions.

Restriction endonucleases ("RE") generally bind
25 with specificity only to their four to eight bp recognition
sites, cleaving the DNA preferably with an at least 2 bp
overhang. Although it is preferable that REs used produce
overhangs of known sequence and characteristic of the
particular RE, other REs, such as those known as class IIS
30 restriction enzymes, which produce overhangs of unknown
sequence can be used to extend initial target subsequences
into longer effective target subsequences. Phasing primers
can also be used to recognize longer effective targer
subsequences. Overhangs of the initial REs can be
35 specifically recognized by hybridization of an adapter
followed by ligation of one strand of this adapter, the

amplification primer. The ligase enzymes, which are used in
this alternativ embodiment of this invention to ligate the
amplification primer, are highly specific in their
hybridization requirements; even one bp mismatch near the
5 ligation site will prevent ligation (U.S. Patent 5,366,877,
Nov. 22, 1994, to Keith et al.; U.S. Patent 5,093,245, Mar.
3, 1992, to Keith et al.). On the other hand, PCR and the
preferred Taq polymerase used therein tolerates hybridization
mis-matches of elongation primers. Thus, PCR embodiments can
10 generate false positive signals which arise from mis-matches
in the hybridization of the oligomer probes to the target
subsequences. However, the PCR embodiments are more flexible
since any desired subsequence can be a target subsequence.
The RE embodiment is limited to the recognition sequences of
15 acceptable REs. However, more than 150 to 200 REs are now
commercially available recognizing a wide variety of
nucleotide sequences.

QEA™ experiments are also adaptable to distinguish
sample sequences into small sets, typically comprising 2 to
20 10 sequences. Such coarser grain analysis requires fewer
subsequence pairs, fewer recognition reactions, and less
analysis time. Alternatively, smaller numbers of target
subsequence pairs can be optimally chosen to distinguish
individually a specific set of sequences of interest from all
25 the other sequences in a sample. These target subsequences
can be chosen either from REs that produce fragments from the
specific sample sequences or, in the case of the PCR
embodiment, from a set of subsequences optimized for this
specific set of sequences.

30   Detailed descriptions of exemplary implementations
for practicing QEA™ recognition reactions and the computer
implemented experimental analysis and design methods are
presented in the following subsections. Detailed
experimental protocols appear in Sec. 6. These
35 implementations are illustrative and not limiting, as this
mbodim nt of th invention may be practiced by any m thod
generating the previously d scribed QEA™ signals.

5.2. RE EMBODIMENTS OF QEA™

The pref rred restriction endonuclease ("RE")
embodiments of QEA™ use novel simultaneous RE and ligase
enzymatic reactions, known as recognition reactions, for
5 generating labeled fragments of the sample sequences to be
analyzed. These labeled fragments are then optionally
amplified by an amplification means, separated according to
length by a separation means, and detected by a detection
means to yield QEA™ signals comprising the identity of the
10 REs cutting each fragment together with each fragment's
length. The RE/ligase subsequence recognition reactions can
specifically and reproducibly generate QEA™ signals with good
signal to noise ratios. Preferred protocols for this
reaction perform all steps, including amplification, in a
15 single tube without any intermediate extractions or buffer
exchanges. This protocol is preferably automatically
performed by standard laboratory robots.

REs bind with specificity to short DNA target
subsequences, usually 4 to 8 bp long, that are termed
20 "recognition sites" and are characteristic of each RE. REs
that are used cut the sequence at (or near) these recognition
sites preferably producing characteristic ("sticky") ends
with single-stranded overhangs, which usually incorporate
part of the recognition site. Type IIS REs, which cut
25 outside of their recognition site, can be used to extend the
initial target subsequence to a longer effective target
subsequence for use in the computer implemented database
lookup.

Preferred REs have a 6 bp recognition site and
30 generate a 4 bp 5' overhang. Less preferred REs generate a 2
bp 5' overhang. These are less preferred since 2 bp
overhangs have a lower ligase substrate activity than 4 bp
overhangs. All RE embodiments can be adapted to 3' overhangs
of two and four bp. In order that an amplification primer
35 hybridization site can be presented on each of the two
strands of the product of the RE/ligase recognition reaction,
as is necessary for experimental amplification. REs

- 78 -

generating 5' and 3' overhangs are preferably not used in the
same recognition reaction. Further, preferred REs have the
following additional properties. Their recognition sites and
overhang sequences are preferably such that an amplification
5 primer can be designed whose ligation does to a cut end does
not recreate the recognition site. They preferably have
sufficient activity below 37°C, and particularly at 16°C, the
optimal ligase temperature, to cut unwanted ligation
products, and are heat inactivated at 65°C and above so that
10 PCR amplification can be performed by simply adding PCR
reagents to the RE/ligase reaction mix. They preferably have
low non-specific cutting and nuclease activities and cut to
completion. The REs selected for a particular experiment
preferably have recognition sites meeting the previously
15 described occurrence and independence criteria. Preferred
pairs of REs for analyzing human and mouse cDNA are listed in
Sec. 6.10.

        Only cDNA fragments with definite and reproducible
lengths dependent only on the source cDNA sequence and
20 independent of cDNA synthesis conditions are of interest.
Only such fragments of definite length are adaptable to the
experimental analysis methods in order to determine their
originating sample sequence. cDNA fragments doubly cut on
each end and by REs have a length dependent only on the
25 sequence of the originating cDNA and are, therefore, of
interest. cDNA fragments singly cut on their 5' end by an RE
and terminated on their 3' end by the poly(A) tail have a
variable and non-reproducible lengths that depend strongly on
cDNA synthesis conditions. Such fragments singly cut on one
30 end by an RE and with a variable length tail on the other are
not of interest. To separate signals from doubly cut
fragments from the unwanted signals from singly cut
fragments, certain RE embodiments of QEA™ exponentially
amplify doubly cut fragments, while only linearly amplifying
35 singly cut fragments. This amplification is preferably done
by the PCR m thod. Other RE embodiments separate singly and
doubly cut fragm nts with a removal means targeted at either

type of fragm nt. Th preferred r moval means comprises a
biotin capture moiety and a streptavidin binding partner.
The removal means can either supplement or replace
differential amplification. On the other hand, cDNA
5 fragments singly cut on their 3' end by an RE and terminated
on their 5' end by a sequence in a fixed relation to the 5'
cap of the source mRNA also have definite lengths and are of
interest. Such fragments can be generated according to a
method herein called 5'-QEA™, which comprises synthesizing
10 cDNA according to the protocol of Sec. 6.3.3, performing
recognition reactions, and separating the fragments of
interest by a removal means. Alternatively, fragments are
also cf interest if they have a definite, sequence dependent
length by being singly cut on their 5' end and by being
15 terminated in a fixed relation with respect to the beginning
of the 3' poly(A)+ tail.

        This invention is adaptable to alternative
amplification means known in the art. If a removal means for
unwanted singly cut fragments is not utilized, alternative
20 amplification means must preferentially amplify doubly cut
fragments with respect to singly cut fragments, in order that
signals from singly cut fragments be relatively suppressed.
On the other hand, if a removal means for singly cut
fragments is utilized in an embodiment, then alternative
25 amplification means can less preferably have no amplification
preference. In RE embodiments using a removal means, this
means can be used either to remove the singly or the doubly
cut fragments. Known alternative amplification means are
listed in Kricka et al., 1995, Molecular Probing, Blotting,
30 and Sequencing, chap. 1 and table IX, Academic Press, New
York. Of these alternative means, those employing the T7 RNA
polymerase are preferred.

        Certain other embodiments use a physical removal
means to directly remove unwanted singly cut fragments,
35 preferably before amplification. Singly cut fragment removal
can be accomplished, e.g., by labeling DNA termini with a
capture moiety pri r to digestion, as by synthesizing the

- 80 -

cDNA with biotinylated primers. After digestion, the singly
cut fragments are then removed by contacting the sample with
a binding partner of the capture moiety, affixed to a solid
phase. Alternatively, the doubly cut fragments can be
5 labeled with a capture moiety, as by amplifying the fragments
with primers one of which is labeled with a capture moiety.
The amplification products are contacted with a binding
partner affixed to a solid support, washed, and then
denatured. Thereby, only doubly cut fragments, one end of
10 which is labeled with a capture moiety, are separated.
Alternately, single stranded fragments can be removed by
single stand specific column separation or single strand
specific nucleases.

This invention is applicable to any removal means
15 meeting the following minimal requirements. The removal
means includes a capture moiety and a binding partner. The
capture moiety is capable of conjugation to DNA oligomers
without disruption of hybridization or chain elongation
reactions. The binding partner is capable of attachment to a
20 solid phase support and can bind the capture moiety to such a
support in DNA denaturing conditions. The preferred removal
means is biotin-streptavidin. Other removal means adaptable
to this invention include various haptens, which are removed
by their corresponding antibodies. Exemplary haptens include
25 digoxigenin, DNP, and fluorescein (Holtke et al., 1992,
Sensitive chemiluminescent detection of digoxigenin labeled
nucleic acids: a fast and simple protocol for applications,
Biotechniques 12(1):104-113 and Olesen et al., 1993,
Chemiluminescent DNA sequencing with multiple labeling,
30 *Biotechniques* 15(3):480-485).

RE/ligase embodiments of QEA™ use recognition
moieties. In any one recognition reaction, each recognition
moiety is capable of hybridizing with and being ligated to
overhangs cut by only one RE. Thereby, the recognition
35 sequ nce of that RE is identified. Recognition moieties
typically comprise partially double stranded DNA oligomers,
each oligomer capable of specifically hybridizing with only

one RE generated sticky end in one recognition reaction. In
the RE/ligase embodim nt using PCR amplification, the
recognition moieties also provide primer means for the PCR
and thereby also provide for labeling and recognition of RE
5 cut ends. For example, using a pair of REs in one
recognition reaction generates doubly cut fragments some with
the recognition sequence of the first RE on both ends, some
with the recognition sequence of the second RE on both ends,
and the remainder with one recognition sequence of each RE on
10 either end. Using more REs generates doubly cut fragments
with all pair-wise combinations of RE cut ends from adjacent
RE recognition sites along the sample sequences. All these
cutting combinations need preferably to be distinguished,
since each provides unique information on the presence of
15 different subsequences pairs, the RE recognition sites,
present in the original cDNA sequence. Thus the recognition
moieties preferably have unique labels which label
specifically each RE cut made in a reaction. As many REs can
be used in a single reaction as labeled recognition moieties
20 are available to uniquely label each RE cut. If the
detectable labeling in a particular system is, for example,
by fluorochromes, then fragments cut with one RE have a
single fluorescent signal from the one fluorochrome
associated with that RE, while fragments cut with two REs
25 have mixed signals, one from the fluorochrome associated with
each RE. Thus all possible pairs of fluorochrome labels are
preferably distinguishable. Alternatively, if certain target
subsequence information is not needed, the recognition
moieties need not be distinctively labeled. In embodiments
30 using PCR amplification, corresponding primers would not be
labeled. If silver staining is used to recognize fragments
separated on an electrophoresis gel, no recognition moiety
need be labeled, as fragments cut by the various RE
combinations are not distinguishable.
35   The recognition reaction conditions are preferably
selected, as d scribed in Sec. 6.4, so that RE cutting and
recognition moiety ligation go to full compl tion: all

recognition sites of all REs in the reacti n are cut and
ligated to a recognition moiety. In this manner, the
fragments generated from a sequence analyzed lie only between
adjacent recognition sites of any RE in that reaction. No
5 fragments remain which include an internal RE recognition
site. Multiple REs can be used in one recognition reaction.
Too many REs in one reaction can cut the sequences too
frequently, generating a compressed length distribution with
many short fragments of lengths between 10 and a few hundred
10 base pairs long that are not clearly resolvable by the
separation means. For example, for gel electrophoresis, if
the fragments are too close in length, fragments should not
be closer than 3 bp on the average. Too many REs also can
generate fragments of the same length and end subsequences
15 from different sample sequences. Finally, where fragment
labels are to be distinguished, no more REs can be used than
can have distinguishably labeled sticky ends. These
considerations limit the number of REs optimally useable in
one recognition reaction. Preferably two REs are used, with
20 one, three and four REs less preferable. Preferable pairs of
REs for the analysis of human cDNA samples are listed in Sec.
6.10.

An additional level of sample sequence
discrimination is possible by detecting occurrences of
25 internal subsequences (here called "third target
subsequences"). The presence or absence of a third internal
subsequences can be used in the computer implemented
experimental analysis methods of this invention along with
identification of the two end subsequences and the fragment
30 length to further discriminate the origin of otherwise
identical fragment signals.

Fragments with specific third internal subsequences
can be detected by either labeling or suppressing such
fragments or with Type IIS REs. To label fragments with a
35 third internal subsequence, probes with distinguishable
labels which bind to this target subsequence ar add d to the
fragments prior to detection, and alternatively prior to

s paration and detection. On detection, fragments with this
third subsequence present will generate a signal, preferably
fluorescent, from the probe. Such a probe could be a labeled
PNA or DNA oligomer. Short DNA oligomers may need to be
5 extended with a universal nucleotide or degenerate sets of
natural nucleotides in order to provide for specific
hybridization. Fragments with a third subsequence can be
suppressed in various manners. The absence of such fragments
is determined by comparing a recognition reaction without the
10 suppressing factors with a reaction with the suppressing
factors. First, in embodiments using PCR amplification, a
probe hybridizing with this third subsequence which prevents
polymerase elongation in PCR can be added prior to
amplification. Then sequences with this subsequence will be
15 at most linearly amplified and their signal thereby
suppressed. Such a probe could be a PNA or modified DNA
oligomer (with the 3' nucleotide being a ddNTP). Second, if
the third subsequence is recognized by an RE, this RE can be
added to the RE-ligase reaction without any corresponding
20 specific primer. Fragments with the third subsequence
thereby have primers on one end only are at most linearly
amplified. Both these embodiments can be extended to
multiple internal sequences by using multiple probes to
recognize the sequences or to disrupt exponential PCR
25 amplification. Type IIS REs which cut a primer close to its
junction with the original cDNA fragment sequence generates
overhangs which are not contiguous with the initial RE
recognition sequence. The sequence of such an overhang can
be used as a third internal subsequence.
30

5.2.1. RECOGNITION MOIETY STRUCTURE

Construction of the recognition moieties, also
herein called adapters or linker-primer oligomers, is
important and is described here in advance of further details
35 of the individual recognition reaction steps. Their basic
structur is first described, followed sec nd by descriptions
of several enhancements adaptable to QEA™ variations. In the

pref rred embodiment, the adapters are partially doubl
stranded DNA ("dsDNA"). Alternatively, the adapters can be
constructed as oligomers of any nucleic acid having
properties corresponding to those of the preferred DNA
5 polymers. In an embodiment employing an alternative
amplification means, the adapters preferably serve as a
primer for that amplification means, if needed.

Turning first to basic adapter structure, Fig. 2A
illustrates the DNA molecules involved in the ligation
10 reaction as conventionally indicated with the 5' ends of the
top strands and the 3' ends of the bottom strands at left.
dsDNA 201 is a fragment of a sample cDNA sequence with an RE
cut at the left end generating, preferably, four bp 5'
overhang 202. Adapter dsDNA 209 is a synthetic substrate
15 provided by this invention. The structure of adapter 209 is
selected to ensure that RE digestion and adapter ligation
preferably go to completion, that generation of unwanted
products and amplification biases are minimized, and that
unique labels are attached to cut ends (if needed). Adapter
20 209 comprises strand 203, called a primer, and a partially
complementary strand 205, called a linker. The primer is
also known as the longer strand of the adapter, and the
linker is also known as the shorter strand of the adapter.

The linker, or shorter strand, links the cDNA cut
25 by an RE to the primer, or longer strand, by hybridizing to
the overhang generated by the RE and to the primer such that
the 3' end of the primer is adjacent to the 5' end of the
overhang. In this configuration, the primer can be
effectively ligated to the cut dsDNA. Therefore, linker 205
30 comprises subsequence 206 complementary to RE overhang 202
and subsequence 207 complementary to 3' end 204 of primer
203. Subsequence 206 is most preferably of the same length
as the RE overhang. Subsequence 207 is preferably eight
nucleotides long, less preferably from 4 to 12 nucleotides
35 long, but can be of any length as long as the linker reliably
hybridizes with nly one primer in any one recognition
reaction at an appropriate T_m. The appropriate T_m should

preferably be less than the self-annealing T_m of primer 203. This ensures that subsequent PCR amplification conditions can be controlled so that linkers present in the reaction mixture will not hybridize and act as PCR primers, and, thereby,

5 generate spurious fragment lengths. The preferable T_m is less than approximately 68°C. Also, linker 205 preferably lacks a 5' terminal phosphate to prevent its ligated to the 3' bottom strand of dsDNA 201. More importantly, lack of a terminal phosphate also prevents self-annealed adapters from

10 ligating and forming dimers. Adapter self-ligation is disadvantageous in that it would compete with adapter ligation to cut cDNA fragments. Further, adapter dimers would be amplified in a subsequent amplification step generating unwanted fragments, termed amplification noise.

15 Terminal phosphates can be removed from linkers using phosphatases known in the art, followed by separation of the enzyme. An exemplary protocol for an alkaline phosphatase reaction is found in Sec. 6.3.4.

Primer, or longer strand, 203 has a 3' end

20 subsequence 204 complementary to 3' end subsequence 2C7 of linker 205. It is preferable that each RE generated overhang is ligated to a unique primer, in each recognition reaction in order that the overhangs generated by each RE can be detected. Consequently, in each recognition reaction primers

25 and linkers are preferably chosen so that each primer is complementary to and hybridizes with only one linker 205 and that each linker which hybridizes with an RE has a unique sequence 207 for hybridizing with a unique primer. In order that the primer/cDNA overhang ligation reaction go to

30 completion, primer 203 preferably does not recreate the recognition sequence of any RE in one recognition reaction when it is ligated with cDNA end 202. Further, primer 203 preferably has no 5' terminal phosphate in order to prevent primer self-ligations. To minimize amplification noise, it

35 is preferred that primer 203 not hybridize with any sequence pr sent in th original sample mixture. If such hybridization occurred, a subsequence PCR step can amplify

unwanted fragments not cut by the initial REs. The T_m of
primer 203 is pr f rably high, in the range from 50° to 80°C,
and more preferably above 68°C. This permits that the
subsequent PCR amplification can be controlled so that only
5 primers and not linkers initiate new chains, the linkers
remaining melted through the PCR cycle. In the case of gel
electrophoretic fragment separation and detection with, e.g.
Ag staining or an intercalating dye, the primer is optionally
unlabeled. For example, this T_m can be achieved by use of a
10 primer having a combination of a G+C content preferably from
40-60%, most preferably from 55-60%, and a length most
preferably 24 nucleotides, and preferably from 18 to 30
nucleotides. Primer 203 is optionally labeled with
fluorochrome 208, although any DNA labeling system that
15 preferably allows multiple labels to be simultaneously
distinguished is usable in this invention. Generally, the
primer, or longer strand, is constructed so that, preferably,
it is highly specific, free of dimers and hairpins, and
capable of forming stable duplexes under the conditions
20 specified, in particular at the desired T_m. Software
packages are available for primer construction according to
these principles, an example being OLIGO™ Version 4.0 For
Macintosh from National Biosciences, Inc. (Plymouth, MN). In
particular, a formula for T_m can be found in the OLIGO™
25 Reference Manual at Eqn. I, page 2.

Fig. 2B illustrates two exemplary adapters and
their component primers and linkers constructed according to
the above description. Adapter 250 is specific for the RE
BamHI, as it has a 3' end complementary to the 5' overhang
30 generated by BamHI. Adapter 251 is similarly specific for
the RE HindIII. Sec. 6.10 contains a more comprehensive,
non-limiting list of adapters that can be used according to
the invention. All synthetic oligonucleotides of this
invention are preferably as short as possible for their
35 functional roles in order to minimize synthesis costs. A
further alternativ illustrated in Fig. 2C is to construct an
adapter by self hybridization of singl stranded DNA in

- 87 -

hairpin loop configuration 212. Subsequences of loop 212 are constructed with similar structure to the corresponding subsequences of linker 205 and primer 203. Exemplary hairpin loop 211 sequences are C_4 to C_{10}.

5   REs generating 3' overhangs are less preferred and require different adapter structures. A preferable basic adapter structure for 3' overhangs is illustrated in Fig. 3A. dsDNA 301 is a fragment of a sample cDNA cut with a RE generating 3' overhang 302. Adapter 309 comprises primer, or longer strand, 304 and linker, or shorter strand, 305.

10 Primer, or longer strand, 304 includes subsequence 306 complementary to and of the same length as 3' overhang 302 and subsequence 307 complementary to linker 305. It also optionally has label 308 which distinctively labels primer 304. As in the case of adapters for 5' overhangs, in order

15 that the RE digestion and ligation reactions go to completion, primer 304 preferably has no 5' terminal phosphate, in order to prevent self-ligations, and preferably has a sequence such that no recognition site for any RE in one recognition reaction is created upon ligation of the

20 primer with dsDNA 301. To minimize amplification, noise, primer 304 should preferably not hybridize with any sequence in the initial sample mixture. The T_m of primer 304 is preferably high, in the range from 50° to 80°C, and more preferably above 68°C. This ensures the subsequent PCR

25 amplification can be controlled so that only primers and not linkers initiate new chains. For example, this T_m can be achieved by using a primer having a G+C content preferably from 40-60%, most preferably from 55-60%, and a primer length most preferably of 24 nucleotide and less preferably of 18-30

30 nucleotides. Each primer 304 in a reaction can optionally have a distinguishable label 308, which is preferably a fluorochrome.

          Linker, or shorter strand, 305 is complementary to

35 and hybridizes with subsequence 307 of primer 304 in a position adjacent t 3' overhang 302. Linker 305 is m st preferably 8 nucleotides long, less pref rably from 4-16

nucleotides, and has no terminal phosphates to prevent self-
ligation. This linker only promotes ligation specificity and
activity and does not link primer 304 to the cut dsDNA, as in
the 5' case. Further, linker 305 T_m should preferably be
5 less than primer 304 self-annealing T_m. This insures that
subsequent PCR amplification conditions can be controlled so
that linkers present in the reaction mixture will not
hybridize and act as PCR primers, and, thereby, generate
spurious fragment lengths. Fig. 3B illustrates an exemplary
10 adapter with its primer and linker for the case of the RE
NlaIII. As in the 5' overhang case, a 3' adapter can also be
constructed from a hairpin loop configuration.

        Next, several adapter structural enhancements are
described. The use of these enhancements is detailed in the
15 subsequent protocol descriptions. In one alternative, the
adapter primer strand can have a conjugated capture moiety in
addition to or in place of a conjugated label moiety. Such a
label moiety is advantageous in separating various classes of
RE/ligase reaction products by binding the capture moiety to
20 its binding partners. Acceptable and preferred capture
moieties and binding partners have been previously described.
Further, when a primer has a conjugated capture moiety,
particularly biotin which form a streptavidin complex that is
difficult to dissociate, it can advantageous to include a
25 release means in the primer in order to achieve controlled
release from the bound capture moiety. Release means can
involve including subsequences in the primer which can be
cleaved in a controlled manner. One exemplary such
subsequence is one or more uracil nucleotides. In this case
30 digestion with uracil DNA glycosylase (UDG) and subsequent
hydrolysis of the sugar backbone at an alkaline pH effects
releases. Another exemplary such subsequence is the
recognition subsequence of an RE which cuts extremely rarely
if at all in the sequences of the sample. A preferred RE of
35 this sort for human cDNA sequences is AscI, which has an 8 bp
recogniti n sequence that rarely, if ever, occurs in
mammalian DNA. AscI is furth r advantageously active at the

ends of DNA molecules.   In this case, digestion with this RE,
i.e., AscI, will release strand 2351.

In another enhancement, adapters can be constructed
from hybrid primers which are designed to facilitate the
5 direct sequencing of a fragment or the direct generation of
RNA probes for in situ hybridization with the tissue of
origin of the DNA sample analyzed.   Hybrid primers for direct
sequencing are constructed by ligating onto the 5' end of
existing primers the M13-21 primer, the M13 reverse primer,
10 or equivalent sequences.   Fragments generated with such
hybrid adapters can be removed from the separation means and
amplified and sequenced with conventional systems.   Such
sequence information can be used both for a previously known
sequence to confirm the sequence determination and for a
15 previously unknown sequence to isolate the putative new gene.
Hybrid primers for direct generation of RNA hybridization
probes are constructed by ligating onto the 5' end of
existing primers the phage T7 promoter.   Fragments generated
with such hybrid adapters can be removed using the separation
20 means and transcribed into anti-sense RNA with conventional
systems.   Such probes can be used for in situ hybridization
with the tissue of origin of the DNA sample to determine in
precisely what cell types a signal of interest is expressed.
Such hybrid adapters are illustrated in Sec. 6.8.
25   In a further enhancement, the previously described
adapters are used but the PCR primers strands have a extra
subsequence 3' to the adapter primer strands in order to act
as phasing primers.   That is the PCR amplification reaction
is used to recognize additional nucleotides beyond the
30 initial RE target recognition subsequence.   Fig. 2D
illustrates such alternative phasing primers.   In that
figure, sample dsDNA 201 is illustrated after blunt-ending
RE/ligase reaction products but just prior to a PCR
amplification cycle.   dsDNA 201 has been cleaved at position
35 221 producing overhang 202 by an RE recognizing target
recognition subsequence 227, has been ligat d to adapter
primer strand 203, and has been completed to a blunt ended

double strand by strand 220 by incubation at 72°C for 10
minutes. For definiteness and without limitation, the RE
recognition subsequence 227 typically extends 1 bp beyond
overhang 202. Other relative positions depend on the lengths
5 of the overhang and the recognition sequence. Alternative
PCR phasing primer 222, illustrated with its 5' end at the
left, comprises subsequence 223, with the same sequence as
strand 203; subsequence 224, with the same sequence as the RE
overhang 202; subsequence 225, with a sequence consisting of
10 a remaining portion of RE recognition subsequence 227, if
any; and subsequence 226 of P nucleotides. Length P is
preferably from 1 to 6 and more preferably either 1 or 2.
Subsequences 223 and 224 hybridize for PCR priming with
corresponding subsequences of dsDNA 201. Subsequence 225
15 hybridizes with any remaining portion of recognition
subsequence 227, typically 1 bp. Subsequence 226 hybridizes
only with fragments 201 having complementary nucleotides in
corresponding positions 228. When P is 1, PCR primer 222
selects for PCR amplification 1 of 4 possible fragments 201;
20 when P is 2, 1 of 16 are selected. Using 4 (or 16) primers
222, each with one of the possible (pairs of) nucleotides, in
4 (16) aliquots or RE/ligase reaction products selects for
amplification one of the possible fragments 201. These
primers are similar to phasing primers (European Patent
25 Application No. O 534 858 A1, published Mar. 31, 1993).

        The effect of using PCR primers 222, having
subsequences 226 of length P bp, is to extend the initially
recognized RE target subsequence into an effective target
subsequence, which is the initial RE target subsequence
30 concatenated to a subsequence complementary to subsequence
226. Thereby, many additional target subsequences can be
recognized while retaining the specificity and exactness
characteristic of the RE embodiment. For example, REs
recognizing 4 bp subsequences can be used in such a combined
35 reaction with an effective 5 or 6 bp target subsequence,
which need not be palindromic. REs rec gnizing 6 bp
sequences can be used in a c mbined reaction to recognize 7

or 8 bp sequences.   Such eff ctive recognition sequences are
input to the computer implemented design and analysis methods
subsequently described.   .

        In a further enhancement, additional subsequence
5 information 'can be generated from adapters comprising primers
with specially placed Type IIS RE recognition subsequence
followed by digestion with the Type IIS RE and sequencing of
the generated overhang.   In a preferred embodiment, the Type
IIS recognition subsequence is placed so that the generated
10 overhang is contiguous with the original recognition
subsequence of the RE that cut the end to which the adapter
hybridizes.   In this embodiment, an effective target
subsequence is formed by concatenating the sequence of the
Type IIS overhang and the original recognition sequence.   In
15 another embodiment, the Type IIS recognition sequence is
placed so that the sequence of the generated overhang is not
contiguous with the original recognition sequence.   Here, the
sequence of the overhang is used as an third internal
subsequence in the fragment.   In both cases, the additionally
20 recognized subsequence is used in the computer implemented
experimental analysis methods to increase the capability of
determining the source sequence of a fragment.   This
enhancement is illustrated in Figs. 17A-E and is described in
detail in Sec. 5.2.3 ("The SEQ-QEA™ Embodiment").   The
25 primers used in the SEQ-QEA™ embodiment advantageously
included combined enhancements, including label moieties,
capture moieties, and release means.

        It will be apparent to those of skill in the art
that the previously described primers and linkers can be
30 enhanced with combinations of the previously described
embodiment and with other alternatives known in the art to
practice further embodiments and refinements of the RE/ligase
embodiment of QEA™.   This invention comprises these
substantially similar variations of the embodiments described
35 herein.

5.2.2. . RE/LIGASE METHOD STEPS

The steps of the preferred RE/ligase embodiment of
QEA™ comprise: first, in one reaction cutting a cDNA sample
with one or more REs, hybridizing adapters corresponding to
the RES, and ligating the primers of the adapters on the cut
5 ends; second, amplifying the cut fragments, if necessary; and
third, separating the fragments according to length and
detecting fragment lengths and fragment target end
subsequences. If necessary, prior to the first step, the
cDNA sample can be synthesized by methods commonly known in
10 the art, such as those described in Sec. 6.3. Optionally,
following the amplification step, additional steps to remove
unwanted DNA fragments or RE/ligase reaction products prior
tc separation detection can increase QEA™ signal to noise
ratio or simplify interpretation of the resulting signals.
15 Additional Re/ligase embodiments are described, including
those known as 5'-QEA™ and SEQ-QEA™.

In more detail, the RE/ligase embodiment can begin
with pre-synthesized cDNA, or with a tissue sample or mRNA
from which cDNA is to be synthesized. When cDNA is to be
20 synthesized, the exemplary methods and procedures of Sec. 6.3
can be used. QEA™ does not require cloning into a vector.
In the case of a tissue sample, a first step is the largely
conventional separation of RNA from the tissue sample.
Separated RNA is preferably poly(A)+ purified RNA, mRNA
25 separated from particular cellular fractions, or less
preferably total cellular RNA. The steps of separation
involve RNase extraction, DNase treatment and mRNA
purification according to protocols, e.g., of Sec. 6.3.1.
First and second strand cDNA synthesis from mRNA can be
30 performed according to the protocols of Sec. 6.3.2, or the
less preferred protocols of Sec. 6.3.4. In the case of small
quantities of mRNA or where it is advantageous to have full-
length cDNA including complementary sequences out to the 5'
cap of the source mRNA, the preferred synthesis protocols of
35 Sec. 6.3.3, or functionally equivalent protocols, can be
us d.

However obtained, it is important that cDNA used in
the RE/ligase mbodiment of QEA™ not have any terminal
phosphates. This is to minimize noise in subsequent fragment
length separation and detection caused by exponential
5 amplification of unwanted fragments singly cut on one end by
an RE and terminated on the other by a variable length
oligo(dT) tail. Significant background noise can arise from
exponential amplification of singly cut fragments whose blunt
ends have ligated to form a single dsDNA with two cut ends
10 having ligated primers, an apparently doubly cut fragment.
The lengths of such fragment vary depending on cDNA synthesis
conditions and produce diffuse background noise on gel
electrophoresis, which obscures sharp bands from the normally
doubly cut fragments. This background can be eliminated by
15 preventing blunt end ligation of such singly cut cDNA
fragments by initially removing all terminal phosphates from
the cDNA sample, without otherwise disrupting the integrity
of the cDNA. Thus, the final preparation step of a DNA
sample is removal of terminal phosphates from the cDNA
20 sample, if needed.

Thus the final preparation step of a cDNA sample is
removal of terminal phosphates, if needed. Terminal
phosphate removal is preferably done with a heat-inactivated
phosphatase. Phosphatase activity is preferably removed
25 prior to RE digestion and adapter ligation step in order to
prevent interference with the intended ligation of adapters
to doubly cut fragments. Heat inactivation allows
phosphatase removal without a separation or extraction step.
A preferred phosphatase comes from cold living Barents Sea
30 (arctic) shrimp (U.S. Biochemical Corp.) ("shrimp alkaline
phosphatase" or "SAP"). Terminal phosphate removal need be
done only once for each population of cDNA being analyzed.
In other embodiments alternative phosphatases can be used for
terminal phosphate removal, such as calf intestinal
35 phosphatase-alkaline from Boehringer Mannheim (Indianapolis,
IN). Those that are not heat inactivated require a step to

separate the phosphatase from the cDNA sample before the RE/ligase reactions, such as by ph nol-chloroform extraction.

The prepared cDNA is then separated into batches of from 1 picogram ("pg") to 200 nanograms ("ng") of cDNA each,
5 and each batch is separately processed by the further steps of the method. A number of batches sufficient for whichever QEA™ mode is to be practiced are made. For a tissue mode experiment, to analyze gene expression, preferably, from a majority of expressed genes in a human tissue, the presence
10 of about 15,000 distinct cDNA sequences needs to be determined. By way of example, one sample is divided into approximately 50 batches, each batch is then subject to an RE/ligase recognition reaction to generate approximately 200-500 fragments, and more preferably 250 to 350 fragments of 10
15 to 1000 bp in length, the majority of fragments preferably having a distinct length and being uniquely derived from one cDNA sequence. A preferable tissue mode analysis entails approximately 50 batches generating approximately 300 bands each. For query mode experiments, fewer recognition
20 reactions are employed since only a subset of the expressed genes are of interest, perhaps approximately from 1 to 100. The number of recognition reactions in an experiment can then number approximately from 1 to 10 and an approximately from 1 to 10 cDNA batches are prepared.
25   Following cDNA preparation is the important step of simultaneous RE cutting of and adapter ligation to the sample cDNA sequences. The prepared sample is cut with one or more REs. The number of REs and associated adapters preferably are limited so that both a compressed length distribution
30 consisting of shorter fragments is avoided and enough distinguishable labels are available for all the REs used. Alternatively, REs can be used without associated adapters in order that the amplified fragments not have the associated recognition sequences. Absence of these sequences can be
35 used to additionally differentiate genes that happen to produce fragments of identical length with particular REs.

In the same reaction mix, herein called the Qlig
mix, REs, adapters and ligase enzyme are simultaneously
present for concurrent adapter ligation and RE cutting. The
amount of RE enzyme in the reaction is preferably
5 approximately a 10 fold unit excess. Substantially greater
quantities are less preferred because they can lead to star
activity (non-specific cutting), while substantially lower
quantities are less preferred because they will result in
less rapid and only partial digestion and hence incomplete
10 and inaccurate characterization of the subsequence
distribution. REs and corresponding adapters are chosen
according to the previous description. Table 10 in Sec. 6.10
lists exemplary REs and corresponding primers and linkers.
Table 11 in Sec. 6.10.1 lists exemplary combinations for
15 biotin labeled primers. The method is adaptable to any
ligase enzyme that is active in the temperature range 10 to
37°C. T4 DNA ligase is the preferred ligase. In other
embodiments, cloned T4 DNA ligase or T4 RNA ligase can also
be used. In a further embodiment, thermostable ligases can
20 be used, such as Ampligase™ Thermostable DNA Ligase from
Epicenpre (Madison, WI), which has a low blunt end ligation
activity. These ligases in conjunction with the repetitive
cycling of the basic thermal profile for the RE-ligase
reaction, described in the following, permit more complete RE
25 cutting and adapter ligation.

Also present in the Qlig mix are necessary buffers,
as known in the art, and ATP. An excess of primers is
preferably present in the Qlig mix in order than subsequent
amplification can be performed in an automated manner.
30 Preferably primers and linkers are present approximately in
the ratio of 20:1 and to an adequate total primer amount of
approximately 20 pm where 1 ng of cDNA is used. Less
preferably the ratio is 10:1. Also, Betaine (Sigma
Chemicals) is preferably present in the Qlig reaction mix.
35 Betaine has been found to improve the uniformity of signals
from fragments that are at approximat ly the same original

concentration by aiding ligati n activity. Betaine also
improves the PCR amplification of hard to amplify products.
            RE/ligase reaction conditions are optimized to
minimize unwanted products. As previously explained,
5 terminal pho'sphate removal from cDNA samples prevents
unwanted ligation of cDNA blunt ends together and subsequent
exponential amplification of the resulting dimers. Another
class of unwanted products are fragment concatamers, formed
when the sticky ends of cut cDNA fragments hybridize and
10 ligate together. Fragment concatamers are removed by
maintaining restriction enzymes activity during ligation in
order to cut any unwanted concatamers. Further, ligated
primers terminate further RE cutting, since primers do not
recreate RE recognition subsequences. A high molar excess of
15 adapters is, therefore, preferable to limit concatamer
formation by driving the RE and ligase reactions toward
complete digestion and adapter ligation. Finally, unwanted
adapter self-ligation is prevented since primers and linkers
lack terminal phosphates (preferably due to synthesis without
20 phosphates or less preferably due to pretreatment thereof
with phosphatases).
            The temperature profile of the RE/ligase reaction
is important for complete cutting and ligation. The
preferred protocol has several steps. The first step is at
25 the optimum RE temperature for a time sufficient to achieve
substantially complete cutting, for example 37 °C for 30
minutes. The ligase used is preferably active during the
first step. The second step is a ramp at -1 °C/min down to
an optimum temperature for adapter annealing and primer
30 ligation, for example, 16 °C. The third step achieves
substantially complete primer ligation of cut products, and
is, for example, at 16 °C for 60 minutes. The REs used are
preferably active during this third step. The fourth step is
again at the temperature for optimum RE activity to achieve
35 complete cutting of recognition sites and unwanted ligation
products, for example at 37 °C for 15 minutes. The fifth
step is to heat inactivate the Qlig enzymes and is, f r

example, above 65 °C. If the PCR amplification is to be
performed immediately, as in the preferred single tube
protocol of Sec. 6.4.1., this fifth step is at 72 °C for 20
minutes and performs additional reactions to be subsequently

5 described. If the PCR amplification is not to be immediately
performed, the Qlig reaction results are held at 4 °C, as in
the much less preferred multi-tube protocol as Sec. 6.4.5.
This temperature profile, together with the subsequence PCR
profile, is illustrated in Fig. 16D.

10   A less preferred profile involves repetitive
cycling of the first four steps of the temperature protocol
described above, that is from an optimum RE temperature to an
optimum annealing and ligation temperature, and back to an
optimum RE temperature. The additional temperature cycles

15 act to further drive the RE/ligase reactions to completion.
With this profile, it is preferred to use thermostable ligase
enzymes. The majority of restriction enzymes are active at
the conventional 16 °C ligation temperature and hence prevent
unwanted ligations without thermal cycling. However,

20 temperature profiles comprising alternating optimum ligation
conditions and optimum RE conditions can cause both enzymatic
reactions to proceed more rapidly than if at one constant
temperature. An exemplary profile comprises periodically
cycling between a 37 °C optimum RE temperature to a 16 °C

25 optimum annealing and ligation temperature at a ramp of
-1 °C/min, then to a 16 °C optimum ligation temperature, and
then back to the 37 °C optimum RE temperature. Following
completion of approximately 2 to 4 of these temperature
cycles, the RE and ligase enzymes are heat inactivated by a

30 final stage above 65 °C for 10 minutes.

These thermal profiles are easily controlled and
automated by the use of commercially available computer
controlled thermocyclers, for example from MJ Research
(Watertown, MA) or Perkin Elmer (Norwalk, CT).

35   The Qlig mix and reaction temperature profile are
designed to achieve the substantially complete cutting of all
RE recognition sites pres nt in the analyzed sequence mixture

and the substantially complete ligation of primers to cut
ends, each primer being unique in one reaction for one
particular RE cut end. The fragments generated are limited
by adjacent RE recognition sites, with substantially no
5 fragments having an internal undigested sites. Further, a
minimum of unwanted self-ligation products and concatamers is
formed. This invention is adaptable to other temperature
profiles which achieve the same effect of substantially
complete cutting and ligation. Exemplary alternative
10 profiles are described in the accompanying examples in Sec.
6.4.

Following the RE/ligase step is a step for
amplifying the doubly cut cDNA fragments. Although PCR
protocols are described in the exemplary embodiment of this
15 invention, any amplification method that selects fragments to
be amplified based on end sequences is adaptable to this
invention (see above). With high enough sensitivity of
detection means, or even single molecule detection means, the
amplification step can be dispensed with entirely. This is
20 preferable as molecular amplification often distorts the
quantitative response of this method.

PCR amplification protocols used in this invention
are designed to have maximum specificity and reproducibility.
First, PCR amplification produces fewer unwanted products if
25 the linkers remain substantially melted and unable to
initiate DNA strands, such as by performing all amplification
steps at a temperature near or above the T_m of the linker.
Second, amplification primers, typically strand 203 of Fig.
2A (and 304 of Fig. 3A), are preferably designed for high
30 amplification specificity by having a high T_m, preferably
above 50 °C and most preferably above 68 °C, to ensure
specific hybridization with a minimum of mismatches. They
are further chosen not to hybridize with any native cDNA
species to be analyzed. The previously described phasing
35 primers, which are alternatively used for PCR amplification,
have similar properties. Third, the PCR temperature profile
is preferably designed f r sp cificity and reproducibility.

- 99 -

High annealing temp ratures minimize primer mis-
hybridizations. Longer extension times reduce PCR bias
related to smaller fragments. Longer melting times reduces
PCR amplification bias related to high G+C content. A
5 preferred PCR temperature cycles is 95 °C for 30 sec., then
57 °C for 1 min., then 72 °C for 2 min. This preferred PCR
temperature profile is illustrated in Fig. 16D. Fourth, it
is preferable to include Betaine in the PCR reaction mix, as
this has been found to improve amplification of hard to
10 amplify products. To further reduce bias, large
amplification volumes and a minimum number of amplification
cycles, typically between 10 and 30 cycles, are preferred.

Any other techniques designed to raise specificity,
yield, or reproducibility of amplification are applicable to
15 this method. For example, one such technique is the use of
7-deaza-2'-dGTP in the PCR reaction in place of dGTP. This
has been shown to increase PCR efficiency for G+C rich
targets (Mutter et al., 1995, *Nuc. Acid Res.* 23:1411-1418).
For a further example, another such technique is the addition
20 of tetramethylammonium chloride to the reaction mixture,
which has the effect of raising the T_m (Chevet et al., 1995,
Nucleic Acids Research 23(16):3343-3344).

It can be advantageous to process multiple
identical samples of RE/ligase reaction products, *e.g.* the
25 processed Qlig mix, with multiple PCR amplifications.
Amplifications of multiple identical samples with the same
number of cycles serves to check reliability and quantitative
response by comparing signals from each of the separately
amplified aliquots. Amplifications of multiple identical
30 samples with an increasing number of amplification cycles,
for example 10, 15, and 20 cycles, are preferable in that
amplifications with a lower number of cycles can detect more
prevalent fragments in a more quantitative manner, while
amplification with a higher number of cycles can detect less
35 prevalent fragments but less quantitatively.

It is preferable to process PCR amplification in
the same reaction tube as th RE/ligase reaction, as this

- 100 -

promotes automation. First, a PCR reaction mix, herein
called the QPCR mix, is made from appropriate DNA
polymerases, dNTPs, and PCR buffer, but without any primer
strands. Exemplary QPCR mix compositions can be found in the
5 examples of Sec. 6.4. The QPCR mix is placed in a reaction
tube, and a layer of wax melting near but below 72 °C is
layered above the QPCR mix. The Qlig mix is placed above the
wax layer and processed according to the previously described
temperature profile, which does not melt the wax. When the
10 RE/ligase reactions are complete, the tube is incubated at
72°C for 20 min. This incubation melts the linkers from the
fragments, melts the wax layer and allows the processed Qlig
mix and the QPCR mix to combine, and finally, permits the DNA
polymerase to complete the fragments to blunt-ended dsDNA.
15 After this incubation, the PCR temperature profile is
performed according to the preferred protocol for a certain
number of cycles.

It is important in the preferred single tube
embodiment that the Qlig and QPCR mixes do not intermingle
20 before the intended step. Even slight mixing due to hairline
cracks in the wax layer can contaminate the reactions. The
preferred wax to prevent such intermingling is a mixture of
Paraffin wax and Chillout™ 14 wax in a 90:10 ratio,
respectively. The paraffin is a highly purified paraffin wax
25 melting between 58 °C and 60 °C such as can be obtained from
Fluka Chamical, Inc. (Ronkonkoma, N.Y.) as Paraffin Wax cat.
no. 76243. Chillout 14 Liquid Wax is a low melting, purified
paraffin oil available from MJ Research. This wax layer is
created in the following manner. The reaction tubes are pre-
30 waxed by melting the preferred wax onto the upper half of the
sides of the tubes. The QPCR mix is added carefully avoiding
this wax layer. Then the wax layer is melted onto the
surface of the QPCR mix by incubating the tubes at 75°C for 2
min. The wax layer is then carefully solidified by
35 decreasing the temperature of the tubes by 5°C every 2 min.
until a final temperature of 25°C is r ached. The Qlig mix
is then gently added on top of this wax surface. This single

tube protocol is adaptable to other less preferable waxes
that melt at approximately at 72°C, such as Ampliwax beads
(Perkin-Elmer, Norwalk, CT). Further, other so called PCR
"hot-start" procedures can be used, such as those employing
5 heat sensitive antibodies (Invitrogen, CA) to initially block
the activity of the polymerase.

Alternatively, PCR amplification can be performed
in a separate tube. In this case the QPCR mix is prepared in
a second tube. The first tube with the processed Qlig mix is
10 incubated at 72°C for approximately 10 min. in order to melt
the linker from the fragments. An aliquot of the Qlig mix is
then combined with the QPCR mix in the second tube, and a
further incubation at 72°C for 10 minutes completes the
fragments to blunt-ended dsDNA. After this incubation, the
15 PCR temperature profile is performed according to the
preferred protocol for a certain number of cycles.

Following the amplification step, optional cleanup
and separation steps prior to length separation and fragment
detection can be advantageous to substantially eliminate
20 certain unwanted DNA strands and thereby to improve the
signal to noise ratio of QEA™ signals, or to substantially
separate the reaction products into various classes and
thereby to simplify interpretation of detected fragment
patterns by removing signal ambiguities. For example, unused
25 primer strands and single strands produced by linear
amplification are unwanted in later steps. These steps are
based on previously described primer enhancements including
conjugated capture moieties and release means.

In one embodiment of these optional steps where one
30 of the two primers used has a conjugated capture moiety, QEA™
reaction products fall into certain categories. These
categories, described without limitation in the case where
the capture moiety is biotin, are:

a)   dsDNA fragments neither strand of which has a biotin
35   moiety;

b)   dsDNA fragments having only one strand with a conjugated
       biotin moi ty;

c)   dsDNA molecule fragments having biotin moieties
      conjugated to both strands; and
d)   unwanted ssDNA strands with and without conjugated
      biotin.

5 The additional method steps comprise contacting the amplified
fragments with streptavidin affixed to a solid support,
preferably streptavidin magnetic beads, washing the beads to
in a non-denaturing wash buffer to remove unbound DNA, and
then resuspending the beads in a denaturing loading buffer
10 and separating the beads from this buffer. The denatured
single strands are then passed to the separation and
detection steps.

       As a results of these steps only the strand of
category "b" without biotin is removed in the loading buffer
15 for separation and detection. Thereby, only fragments cut on
either end by different REs and freed from single stranded
contaminants are separated and detected with minimized noise.
Category "a" products are not bound to the beads and are
washed away in the non-denaturing wash buffer. Similarly,
20 class "d" products without biotin moieties are washed away.
All products with a conjugated biotin are retained by the
streptavidin beads after washing. The denaturing loading
buffer denatures categories "b" and "c" products attached to
the beads, but both strands of category "c" products have
25 conjugated biotin and remain attached to the beads.
Similarly, class "d" products with conjugated biotin are
retained by the beads.

       In another embodiment, the biotinylated primer can
include a release means in order to recover fragments of
30 class "c". After the step of suspension in a denaturing
buffer, the releasing means, e.g. UDG or AscI, can be applied
to release the biotinylated strands for separation and
detection. Fragments detected at this second separation in
addition to those previously detected then represent class
35 "c" products.

       Further embodiments will be apparent to those of
skill in the art. For example, two or more types of capture

moieties can be used in a single reaction to separate
different classes of pr ducts. Capture moieties can be
combined with release means to achieve similar separation.
Label moieties can be combined with capture moieties to
5 verify separations or to run reactions in parallel.

This invention is adapted to other less preferred
means for single strand separation and product concentration
that are known in the art. For example, single strands can
be removed by the use of single strand specific exonucleases.
10 Mung Bean exonuclease, Exo I or S1 nuclease can be used, with
Exo I preferred because of its higher specificity for single
strands while S1 is least preferred. Other methods to remove
unwanted strands include the affinity based methods of gel
filtration and affinity column separation. Amplified
15 products can be concentrated by ethanol precipitation or
column separation.

The last QEA™ step is separation according to
length of the amplified fragments followed by detection the
fragment lengths and end labels (if any). Lengths of the
20 fragments cut from a cDNA sample typically span a range from
a few tens of bp to perhaps 1000 bp. Any separation method
with adequate length resolution, preferably at least to three
base pairs in a 1000 base pair sequence, can be used. It is
preferred to use gel electrophoresis in any adequate
25 configuration known in the art.

Gel electrophoresis is capable of resolving
separate fragments which differ by three or more base pairs
and, with knowledge of average fragment composition and with
correction of composition induced mobility differences, of
30 achieving a length precision down to 1 bp. A preferable
electrophoresis apparatus is an ABI 377 (Applied Biosystems,
Inc.) automated sequencer using the Gene Scan software (ABI)
for analysis. The electrophoresis can be done by suspending
the reaction products in a loading buffer, which can be non-
35 denaturing, in which the dsDNA remains hybridized and carries
the labels (if any) of both primers. The buffer can als be
denaturing, in which the dsDNA separates into single strands

- 104 -

that typically are expected to migrate together (in he
absence of large average differenc s in strand composition or
significant strand secondary structure). The length
distribution is detected with various detection means. If no
5 labels are used, means such as Ag staining and intercalating
dyes can be used. Here, it can be advantageous to separate
reaction products into classes, according to the previously
described protocols, in order that each band can be
unambiguously identified as to its target end subsequences.
10 In the case of fluorochrome labels, since multiple
fluorochrome labels can be typically be resolved from a
single band in a gel, the products of one recognition
reaction with several REs or other recognition means or of
several separate recognition reactions can be analyzed in a
15 single lane. However, where one band reveals signals from
multiple fluorochrome labels, interpretation can be
ambiguous: is such a band due to one fragment cut with
multiple REs or to multiple fragments each cut by one RE. In
this case, it can also be advantageous to separate reaction
20 products into classes.

Preferred protocols for the specific RE embodiments
are described in detail in Sec. 6.4.

5.2.3. THE SEQ-QEA™ EMBODIMENT

25   SEQ-QEA™ is an alternative embodiment of the
preferred method of practicing a RE/ligase embodiment of QEA™
method as previously described in Sec. 5.2.2. By the use of
adapters comprising specially constructed primers bearing a
recognition site for a Type IIS RE, a SEQ-QEA™ method is able
30 to identify an additional 4-6 terminal nucleotides adjacent
to the recognition subsequence of the RE initially cutting a
fragment. Thereby, the effective target subsequence is the
concatenation of the initial RE recognition subsequence and
the additional 4-6 terminal nucleotides, and has, therefore,
35 a length of at least from 8 to 12 nucleotides and preferably
has a length of at least 10 nucleotides. This longer
effective target subsequence is then used in QEA™ analysis

methods as described in Sec. 5.4 ("QEA™ Analysis and Design
Methods") which inv lves searching a database of sequences to
identify the sequence or gene from which the fragment
derived. The longer effective target subsequence increases
5 the capability of these methods to determine a unique source
sequence for a fragment.

In this section, for ease of description and not
limitation, first shall be described Type IIS REs, next the
specially constructed primers, and then the additional method
10 steps of a SEQ-QEA™ method used to recognize the additional
nucleotides.

A Type IIS RE is a restriction endonuclease enzyme
which cuts a dsDNA molecule at locations outside of the
recognition sequence of the Type IIS RE (Szybalski et al.,
15 1991, Gene 100:13-26). Fig. 17C illustrates Type IIS RE 1731
cutting dsDNA 1730 outside of its recognition subsequence
1720 at locations 1708 and 1709. The Type IIS RE preferably
generates an overhang by cutting the two dsDNA strands at
locations differently displaced away on the two strands from
20 the recognition sequence. Although the recognition
subsequence and the displacement(s) to the cutting site(s)
are determined by the RE and are known, the sequence of the
generated overhang is determined by the dsDNA cut, in
particular by its nucleotide sequence outside of the Type IIS
25 recognition region, and is, at first, unknown. Thus in a
SEQ-QEA™ embodiment the overhangs generated by the Type IIS
REs are sequenced. Table 17 in Sec. 6.10.1 lists several
Type IIS REs adaptable for use in the SEQ-QEA™ method and
their relevant characteristics, including their recognition
30 subsequences on both DNA strands and the displacements from
these recognition subsequences to the respective cutting
sites. It is preferable to use REs of high specificity and
generating an overhang of at least 4 bp displaced at least 4
or 5 bp beyond the recognition subsequence in order to span
35 the remaining recognition subsequence of the RE that
initially cut the fragment. FokI and BbvI are most preferred
Type IIS REs for the SEQ-QEA™ method.

Next, the special primers, and the special linkers
if n eded, which hybridize to form the adapters for SEQ-QEA™,
have, in additional to the structure previously described in
Sec. 5.2.1, a Type IIS recognition subsequence whose
5 placement is important in order that the overhang generated
by the Type IIS enzyme be contiguous to the initial target
end subsequence. The placement of this additional
subsequence is described with reference to Figs. 17A-E, which
illustrate steps in a SEQ-QEA™ alternative embodiment. Fig
10 17B schematically illustrates dsDNA 1702, which is a fragment
cut from an original sample sequence on one end by a first RE
and on the other end by a different second initial RE, with
adapters fully hybridized but prior to primer ligation.
Thus, linker strand 1711 has hybridized to primer strand 1712
15 and to the 5' overhang generated by the first RE, and how
fixes primer 1712 adjacent to fragment 1702 for subsequent
ligation. Primer 1712 has recognition subsequence 1720 for
Type IIS RE 1721. Linker 1711, to the extent it overlaps and
hybridizes with recognition subsequence 1720, has
20 complementary recognition subsequence 1721. Additionally,
primer 1712 preferably has a conjugated label moiety 1734,
e.g. a fluorescent FAM moiety. Similarly, linker strand 1713
has hybridized to primer strand 1714 and to the 5' overhang
generated by the second RE. Primer 1714 preferably has a
25 conjugated capture moiety 1732, e.g. a biotin moiety, and a
release means represented by subsequence 1723.

Subsequence 1704 terminating at nucleotide 1707 in
Fig. 17B is the portion of the recognition subsequence of the
first RE remaining after its cutting of the original sample
30 sequence. The placement of the Type IIS RE recognition
subsequence is determined by the length of this subsequence.
Fig. 17A schematically illustrates how the length of
subsequence 1704 is determined by properties of the first RE.
The first initial RE is chosen to be of a type that
35 recognizes subsequence 1703, terminating with nucleotide
1707, of sampl dsDNA 1701, and that cuts the two strands of
dsDNA 1701 at locations 1705 that are located within

- 107 -

recognition subs qu nce 1703. In order that the first RE
recognize a known target subsequence, it is highly preferable
that subsequence 1703 be entirely determined by the first RE
and be without indeterminate nucleotides. As a result of
5 this cutting, overhang subsequence 1706 is generated and has
a known sequence, since it is entirely within the determined
recognition subsequence 1703. Thereby, subsequence 1704, the
portion of the recognition subsequence 1703 remaining on a
fragment cut by the first RE, has a length not less than the
10 length of overhang 1706 and is typically longer. Typically
and preferably, subsequence 1703 is of length 6 and is
palindromic; locations 1705 are symmetrically placed in
subsequence 1703; and overhang 1706 is of length 4.
Therefore, the typical length of the remaining portion 1704
15 of the recognition subsequence 1703 is of length 5.

        The preferred placement of Type IIS recognition
sequence 1720 is now be described with reference to Fig. 17C,
which schematically illustrates dsDNA 1730, which derives
from dsDNA 1702 of Fig. 17B after the further steps cf primer
20 ligation, PCR amplification with primers 1712 and 1714,
binding of capture moiety 1732 to binding partner 1733
affixed to a solid-phase substrate, and binding of Type IIS
RE 1731 to its recognition subsequence 1720. Subsequence
1722 is the subsequence between recognition subsequence 1720
25 and the end of primer 1712 at location 1705. Type IIS RE is
illustrated cutting dsDNA 1730 at nucleotide locations 1708
and 1709 and, thereby, generating an exemplary 5' overhang
1724 between these locations. For this overhang to be
contiguous with the remaining portion 1704 of initial target
30 end subsequence 1703, nucleotide 1709 is adjacent to
nucleotide 1707 terminating subsequence 1704. Therefore,
Type IIS recognition sequence 1720 is preferably placed on
primer 1712 such that the length of subsequence 1704 plus the
length of subsequence 1722 equals the distance of closest
35 cutting of Type IIS RE 1731. For example, in the case of
FokI, since the closest cutting distance is 9 and the typical
length of subsequence 1704 is 5, its r cognition sequence is

- 108 -

preferably placed 5 bp from the end of primer 1712. In the
case of BbvI, since the closest cutting distance is 8, its
recognition sequence is preferably placed 3 bp from the end
of primer 1712.

5   Finally, Fig. 17D schematically illustrates dsDNA
1730 after cutting by Type IIS RE 1731. dsDNA has 5'
overhang 1724 between and including nucleotides 1708 and
1709, where the Type IIS RE cut dsDNA 1730 of Fig. 17C. This
overhang is contiguous with former subsequence 1704, the
10 remaining portion of the recognition sequence of the first
RE, which has been cut off. The shorter strand has primer
1714 including release means represented by subsequence 1723.
dsDNA 1730 remains bound to the solid-phase support through
capture moiety 1732 and binding partner 1724. The absence of
15 label moiety 1734 can be used to monitor the completeness of
cutting by Type IIS RE 1731.

          This invention is also adaptable to other less
preferable placements of recognition sequence 1720. If
recognition sequence 1720 is placed closer to the 3' end of
20 primer 1712 than the optimal and preferable distance, the
overhang produced by Type IIS RE 1731 is not contiguous with
recognition subsequence 1703 of the first RE, and a
contiguous effective target subsequence is not generated. In
this case, optionally, the determined sequence of the Type
25 IIS RE generated overhang can be used as third internal
subsequence information in QEA™ experimental analysis methods
in order to further resolve the source sequence of fragment
1702, if necessary. If recognition sequence 1720 is placed
further from the 3' end of the cut primer than the optimal
30 and preferable distance, the overhang produced by Type IIS RE
overlaps with recognition subsequence 1703 of the first RE.
In this case, the length of the now contiguous effective
target subsequence is less than the sum of the lengths of the
Type IIS overhang and the first RE recognition subsequence.
35 Effective target end subsequence information is, thereby,
lost. In cas recognition sequence 1710 is placed further

fr m the 3' end than th distance of furthest cutting, no
additional information is obtained.

Primer 1714 also has certain additional structure.
First, primer 1714 has capture moiety 1732 conjugated near or
5 to its 5' end. Biotin/streptavidin are the preferred capture
moiety/binding partner pair, which are used in the following
description without limitation to this invention. Second,
primer 1714 has release means represented as subsequence
1723. As previously described, the release means allows
10 controlled release of strand 1735 of Fig. 17D from the
capture moiety/binding partner complex. This alternative is
adaptable to any such controlled release means, including the
cases where subsequence 1723 is one or more uracil
nucleotides and where it is the recognition subsequence of an
15 RE which cuts extremely rarely if at all in the sequences of
the sample. e.g. AscI. Release means are particularly useful
in the case of biotin-streptavidin, which form a complex that
is difficult to dissociate.

Table 18 of Sec. 6.10.1 lists exemplary primers,
20 linkers, and associated REs, for the preferred implementation
cf SEQ-QEA™ in which contiguous effective target end
subsequences are formed. This description has illustrated
the generation of a 5' Type IIS generated overhang. Primers
can equally be constructed to generate a less preferable 3'
25 overhang by using a Type IIS whose closest cutting distance
is on the 3' strand, rather than on the 5' strand.

Finally, the method steps of SEQ-QEA™ are now
described. SEQ-QEA™ comprises, first, practicing the
RE/ligase embodiment of QEA™ using the special primers and
30 linkers previously described followed, second, by certain
additional steps unique to SEQ-QEA™. Figs. 17B-E illustrate
various steps in a SEQ-QEA™ method. Fig. 17B illustrates a
fragment from a sample sequence digested by two different REs
and just prior to primer ligation. Fig. 17C illustrates a
35 sample sequence after primer ligation, chain blunt-ending,
and PCR amplificati n. Thes QEA™ steps are preferably
performed according to the embodiments described in Sec.

5.2.2, but can alternatively be performed by any RE/ligase
embodiment. The additi nal steps uniqu to SEQ-QEA™ include,
first, binding the amplified fragments to a solid-phase
support, also illustrated in Fig. 17C, second, washing the
5 bound fragments, and third, digesting the bound fragments by
the Type IIS RE corresponding to primer 1712 used. The Type
IIS digestion is preferably performed with reaction
conditions suitable to achieve complete digestion, which can
be checked by insuring the absence of optional label moiety
10 1734 after washing the bound, digested sequences. Fig. 17D
illustrates dsDNA fragments 1730 remaining after complete
digestion by the Type IIS RE. Before Type IIS digestion, an
aliquot of the bound, amplified RE/ligase reaction products
is denatured and the supernatant, containing the labeled 5'
15 strands, are separated according to length by, e.g., gel
electrophoresis, in order to determine the length of each
fragment dcubly cut by different REs.

        The subsequent additional SEQ-QEA™ step is
sequencing of overhang 1724. This can be done in any manner
20 known in the art. In a preferred embodiment suitable for
lower fragment quantities, an alternative, herein called a
phasing QEA™ method, can be used to sequence this overhang.
Phasing QEA™ depends on the precise sequence specificity with
which RE/ligase reactions recognize short overhangs, in this
25 case the Type IIS generated overhang. Fig. 17E illustrates a
first step of this embodiment in which a QEA™ method adapter,
which is comprised of primer 1751 with label moiety 1753 and
linker 1750, has hybridized to overhang 1724 in Type IIS
digested fragment 1730 bound to a solid-phase support. By
30 way of example only, overhang 1724 is here illustrated as
being 4 bp long. In this embodiment, special phasing linkers
are used. For each nucleotide position of overhang 1724,
e.g. position 1754, 4 pools of linkers 1750 are prepared.
All linkers in each pool have one fixed nucleotide, i.e. one
35 of either A, T, C, or G, at that position, e.g. position
1755, while random nucle tides in all combinations ar
present at the other three positions. For each nucleotid

- 111 -

position of the overhang, four RE/ligase reactions are
performed according to QEA™ protocols, one reaction using
linkers from one of the four corresponding pools. Linkers
from only one pool, that having a nucleotide complementary to
5 overhang 1724 at position 1754, hybridize without error, and
only these linkers can cause ligation of primer 1751 to the
5' strand of fragment 1730. When the results of the four
RE/ligase reactions are denatured and separated according to
length, only one reaction of the four can produce labeled
10 products at a length corresponding to the length of fragment
1730, namely the reaction with linkers complementary to
position 1754 of overhang 1724. Thereby, by performing four
RE/ligase reactions for each nucleotide position of overhang
1724, this overhang can be sequenced. Optionally, the
15 products of these four RE/ligase reactions can be further PCR
amplified. In a further option, if linkers 1750 comprise
subsequence 1756 that is uniquely related to the fixed
nucleotide in subsequence 1752 and if four separately and
distinguishably labeled primers 1751 complementary to these
20 unique subsequences are used, all four RE/ligase reactions
for one overhang position can be simultaneously performed in
one reaction tube. With this overhang sequencing alternative
embodiment, release means 1723 can be omitted from primer
1714.

25   In an alternate embodiment, sequencing of a 5'
overhang can be done by standard Sanger reactions. Thus
strand 1735 is elongated by a DNA polymerase in the presence
of labeled ddNTPs at a relatively high concentration to dNTPs
in order to achieve frequent incorporation in the short 4-6
30 bp elongation. Partially elongated strands 1735 are released
by denaturing fragment 1730, washing, and then by causing
release means 1723 to release strands 1735 from the capture
moiety bound to the solid phase support. The released,
partially elongated strands are then separated by length,
35 e.g., by gel electrophoresis, and the chain terminating ddNTP
is observed at the length previously observed for that

fragment. In this manner, the 4-6 bp overhang 1724 of each fragment can be quickly sequenced.

The effective target subsequence information, formed by concatenating the sequence of the Type IIS overhang
5   to the sequence of the recognition subsequence of the first RE, is then input into QEA™ Experimental Analysis methods, and is used as a longer target subsequence in order to determined the source of the fragment in question. This longer effective target subsequence information preferably
10 permits exact and unique sample sequence identification.

5.2.4. 5'-QEA™ ALTERNATIVE RE EMBODIMENT

In QEA™ embodiments of this invention, it is important that the one or more fragments of a nucleic acid
15 from a sample which are generated by the recognition reactions be of definite length, that is that the length of each fragment depends only on the sequence of the nucleic acid and not on experimental conditions, e.g., the synthesis conditions of the nucleic acid. Further, it is important for
20 the experimental analysis and design methods of Sec. 5.4 that the length of a fragment be precisely predicable from the nucleotide sequence of the sample nucleic acid. In the preferred RE/ligase embodiments of QEA™, these goals are accomplished primarily by selecting signals from fragments
25 doubly cut on both ends by one or more REs. The nucleotide distance between adjacent RE recognition subsequences is determined only by the sequence of nucleic acid from the sample. Also the described alternatives and extensions generate additional signal information dependent only on the
30 nucleic acid sequence. In these embodiments, nucleic acid, e.g. cDNA, synthesis conditions are then only of indirect importance, in that they preferably adequately represent input mRNA.

Other RE/ligase embodiments utilize signals from
35 fragments of a nucleic acid that, although only singly cut by an RE on one nd, n v rth less have a definite length, dependent only on nucleotide sequence, because of particular

- 113 -

cDNA synthesis conditions that fix the other end. For these
embodi mnts, therefor, the cDNA synthesis conditi ns are of
direct importance, in that these embodiments can only be used
with cDNA synthesized according to the particular conditions.
5 In general, these conditions insure that the cDNA begins or
ends in a known relation, herein called "anchored," to
general landmarks on the input mRNA. In particular,
preferable anchoring landmarks include the 5' end of the
poly(A)+ tail present on the 3' end of the input mRNA, or the
10 cap on the 5' end of the input mRNA. For example, cDNA
fragments terminated on their 5' end in a fixed relation to
the 5' cap of the source mRNA and cut on their 3' end at the
nearest recognition subsequence of a single RE have a
definite length and generate QEA™ signals that can be used to
15 determine the source nucleic acid in the sample. Similarly,
cDNA fragments terminated on their 3' end in a fixed relation
to the 5' end of the poly(A)+ tail present on the source mRNA
and cut on their 5' end at the nearest recognition sequence
of a single RE also have a definite length and generate QEA™
20 signals that can also be used to determine the source nucleic
acid in the sample.

        Turning first the case of 5' anchored cDNA, such
cDNA can be synthesized by a protocol which requires the
presence of an intact 5' cap on the input mRNA. One such
25 exemplary preferred protocol is described in Sec. 6.3.3.
This protocol depends upon using a RNA ligase to ligate to a
source mRNA at the nucleotide adjacent to the 5' cap a DNA-
RNA chimera comprising a first DNA subsequence 5' to the
ribonucleotide triplet GGA at the 3' end of the chimera. The
30 RNA component of the DNA-RNA chimera is preferably GGA, but
any RNA subsequence can be used that promotes effective
ligation by the ligase chosen of the chimera to the source
mRNA. The DNA oligonucleotide component is later used as a
primer and is herein called a "5'-cap-primer"
35 oligonucleotide. This ligation is accomplished by
dephosphorylating input mRNA with an alkaline phosphatase and
then cleaving the 5' cap with an acid pyrophosphatase,

preferably tobacco acid pyrophosphatas , leaving a 5'
phosphate need d for ligati n only on mRNAs having a 5' cap.
During the ligation step, an excess of primer is used to
prevent self-ligations of the input mRNA. The preferred RNA
5 ligase is T4 RNA ligase. First strand synthesis is then
performed with a first DNA primer comprising the first DNA
subsequence. Thereby, all cDNAs originate from input mRNAs
having their 5' cap. Second strand synthesis is then
performed with such second strand primers as are known in the
10 art. Preferabl,y second strand primers are three second
strand primers mixed or in separate pools, each of which
comprises a second DNA subsequence 5' to one of three
oligo(dT) one-nucleotide phasing primers, as known in the art
(Liang et al., 1994, *Nuc. Acid Res.* 22:5763-5764).
15 Alternatively, other primers known in the art could be used,
including, a single oilgo(dT) primer, a sequence specific
primer, or random primers. For small amounts of input mRNA,
the first DNA primer and a second DNA primer comprising the
second DNA subsequence can be used in a PCR reaction to
20 amplify the synthesized cDNA. This QEA™ embodiment is
adaptable to other methods known in the art to produce cDNAs
with a 5' end anchored in a fixed relation to the 5' mRNA
cap, for example the CapFinder™ PCR cDNA Library Construction
Kit Clonetech (Palo Alto, CA). See also Schmidt et al.,
25 1996, *Nuc. Acids. Res.* 24:1789-1791.

        The first and second DNA primer sequences are
preferably chosen according to certain guidelines. First,
they are chosen not to generate by themselves any PCR
products from the cDNA sample nucleic acids. Second, they
30 are of a sufficient length and average base content
(approximately 60% G+C) to hybridize in high stringency
conditions. Third, they have no significant secondary
structure. Finally, they can have included RE recognition
sites, initiators, etc. to promote later cloning or
35 expression. Exemplary first and second primers are described
in Sec. 6.3.3. Software packages are available for primer
construction according to such guidelines, an example being

- 115 -

OLIGO™ Version 4.0 For Macintosh from National Biosciences,
Inc. (Plymouth, MN).

      Having cDNA synthesized according to the exemplary
5' anchoring protocol, the 5'-QEA™ embodiment is performed
5 according to the general methods Sec. 5.2.2, including the
optional cleanup and separation steps. In particular, the
QPCR mix is prepared as previously described. The Qlig mix
includes the one RE chosen to cut the fragment and an
associated adapter with primer excess. These primers are
10 preferably be labeled are most preferably do not have a
conjugated capture moiety. Also included in the Qlig mix in
a quantity sufficient for PCR amplification is an extra
primer, which is the first DNA primer, that is the DNA
portion of the chimera now appearing on the 5' end of the
15 synthesized cDNA, together with a conjugated biotin moiety or
other capture moiety,. The RE/ligase reactions and the
subsequent PCR amplification are performed as previously
described and result in the following classes of fragments.
First, there are fragments singly cut by the chosen RE which
20 are exponentially amplified because of the presence of the
first DNA primer and which have on their 5' ends the biotin
labeled first DNA primer. Second, there are exponentially
amplified fragments doubly cut by the chosen RE which have no
biotin labels. Third, there can be linearly amplified, non-
25 labeled, singly cut fragments. After contacting these
reaction products with streptavidin beads and washing, only
the first class of fragments is retained, that is fragments
singly cut adjacent to the 5' end. Upon resuspending the
beads in a denaturing loading buffer, only the denatured
30 single strands from such fragments generate signals after the
separation and detection steps. These signals have a
definite length, because the RE recognition site nearest the
5' end is determined only by the sequence of the nucleic
acid.

35   Turning to the less preferred case of 3' anchored
cDNA, such cDNA can be synth siz d by protocols known in the
art which utilize phasing prim rs. Such phasing primers can

comprise a first DNA subsequence, which is construct d
according to the previously describ d primer guidelines, 5'
to one of three oligo(dT) one nucleotide phasing primer
subsequences (Liang et al. 1994). Sequences MBTA, MBTC, and
5 MBTG of Sec. 6.3.3 are exemplary of such primers. The
RE/ligase and PCR amplification reactions are carried out
according to the protocol of the 5'-QEA™ embodiment with the
exception that the extra primer used in the Qlig mix is the
first DNA subsequence used in the prior cDNA synthesis with a
10 conjugated biotin or other capture moiety. After completion
of the protocol, signals are only generated from fragments
cut by the chosen RE adjacent to the 3' end. These signals
have a definite length, because the RE recognition site
nearest the 3' end is determined only by the sequence of the
15 nucleic acid.

        The signals generated from the singly cut fragments
according to the protocols of this section can be used in the
computer implemented experimental analysis methods of Sec.
5.4 in order to determine the sample nucleic source of a
20 particular signal. The analysis methods need minimal
adaptation in a manner that will be apparent to one of skill
in the computer arts in order that the 5' or 3' end cDNA
sequence is one of the target end sequences. This adaptation
can be done in several ways, including simply specially
25 marking in the signals that one target end subsequence is the
3' or 5' end as needed or by including in the generated
signal an artificial and not naturally occurring target
subsequence that represents the 3' or the 5' end as
appropriate and concatenating these artificial subsequences
30 to nucleic acid sequences input from a database prior to
computer processing. Similar minimal adaptations to the
computer implemented experimental design methods can be made
in order to create and optimize experiments generating singly
cut fragments.
35   The embodiments described in this section, in
particular 5'-QEA™, can b practiced in combination with QEA™
embodiments h r in described. It will be appar nt t on   f

skill in the art how such combinations can be performed.
Specifically, it is advantageous to combine 5'-QEA™ with SEQ-
QEA™ to obtain signals which include longer effective target
subsequence information on the singly cut end along with
5 information on the distance of the effective target
subsequence from the end of the cDNA.

5.2.5. FURTHER ALTERNATIVE RE EMBODIMENTS

The embodiments of this section remove unwanted
10 RE/ligase reaction products at least partially by utilizing
cDNA with conjugated capture moieties, obtained perhaps from
either first and second strand synthesis with primers having
conjugated capture moieties or from PCR amplification of cDNA
with such primers. The preferred capture moiety is biotin
15 for which the corresponding binding partner is streptavidin
attached to a solid support, preferably magnetic beads.
These embodiments are adaptable to other capture moieties and
corresponding binding partners.

A first QEA™ embodiment in conjunction with
20 sufficiently sensitive detection means can advantageously
minimize or eliminate altogether the PCR amplification step.
PCR amplification disadvantageously has a non-linear response
well known in the arts, depending on such factors as fragment
length, average base composition, and secondary structure.
25 To improve quantitative response, it is preferred to
eliminate the PCR amplification step or at least to minimize
the number of PCR cycles. Then output signal intensity is
more nearly linearly responsive to the abundance of the input
nucleic acids generating that signal.
30   In the previously described RE/ligase embodiments
the amplification step serves both to amplify the signals
from fragments of interest and simultaneously to dilute the
signals from unwanted fragments without a definite sequence-
dependent length and. For example, in the protocol of Sec.
35 5.2.2, fragments doubly cut with REs and ligated to adapters
are exponentially amplified, while unwant d fragments singly
cut by an RE are at best lin arly amplified. After ten

- 118 -

cycles of amplification, since doubly cut fragments are
amplifi d 1000X while singly cut fragments are amplified 10X,
fragments from sample nucleic acids with a relative abundance
of 1% or more can be detected above the background noise
5 while fragments from sample nucleic acids with a relative
abundance of 1% or less can be lost in the unwanted
background. More amplification cycles permit both greater
sensitivity and greater ability to observe rare fragments
from rare sequences.

10   More sensitive detection means decrease the need
for amplification in order to generate observable signals.
In the case of standard fluorescent detection means, a
minimum of 6 x 10^{-18} moles of fluorochrome (approximately 10^5
molecules) is required for detection. Since one gram of cDNA
15 contains about 10^{-6} moles of transcripts, it is possible to
detect transcripts to at least a 1% relative level from
microgram quantities of mRNA. With greater mRNA quantities,
proportionately rarer transcripts are detectable. Labeling
and detection schemes of increased sensitivity permit use of
20 less mRNA. Such a scheme of increased sensitivity is
described in Ju et al., 1995, Fluorescent energy transfer
dye-labeled primers for DNA sequencing and analysis, Proc.
Natl. Acad. Sci. USA 92:4347-4351. Possible single molecule
detection means are about 10^5 times more sensitive than
25 existing fluorescent means (Eigen et al., 1994, *Proc. Natl.*
Acad. Sci. USA 91:5740-5747).

          To minimize or eliminate amplification steps, the
first embodiment described in this section minimizes the need
for amplification in order to dilute unwanted signals by
30 using a capture moiety to remove unwanted singly cut
fragments from the doubly cut fragments of interest. In the
protocols of Sec. 5.2.2, only the doubly cut fragments have
definite lengths dependent only on the sequences of the input
nucleic acids. Singly cut fragments have non-diagnostic
35 lengths depending also on cDNA synthesis conditions. In this
protocol, PCR amplification can be optionally employed to
gen rate suffici nt signal int nsity for detection. It is

not n eded to minimize the background noise g nerated in the
previously d scribed protocols. The steps of this protocol
comprise synthesis of cDNA using a primer labeled with a
capture moiety, circularization of the cDNA, cutting with
5 REs, and ligation to adapters. Singly cut ends are then
removed by contacting the reaction products with a solid
phase to which the binding partner of the capture moiety is
affixed.

        Figs 4A, 4B, and 4C illustrate this alternative
10 protocol, which preferably uses biotin as a capture moiety
for direct removal of the singly cut 3' and 5' cDNA ends from
the RE/ligase reaction products. cDNA first strands are
synthesized according to the method of Sec. 6.3.3 using, for
example, an oligo(dT) primer with a biotin molecule linked to
15 a thymidine nucleotide. For example, such a primer is
T_nT(biotin)T_m, with n approximately equal to m, and with n + m
sufficiently large, approximately 12 to 20, so that the
primer will reliably hybridize to the poly(A) tail of mRNA.
Other biotin labeled primers may also be used, such as random
20 hexamers. Double stranded cDNA is then synthesized, also
according to Sec. 6.3.3. In this embodiment, terminal
phosphates are retained. Fig. 4A illustrates such a cDNA 401
with ends 407 and 408, poly(dA) subsequence 402, oligo(dT)
primer 403 with biotin 404 attached. Subsequence 405 is the
25 recognition sequence for RE_1; subsequence 406 is the
recognition sequence for RE_2. Fragment 409 is the cDNA
sequence defined by these adjacent RE recognition sequences.
Fragments 423 and 424 are singly cut fragments resulting from
RE cleavages at subsequences 405 and 406.
30   Next, the cDNA is ligated into a circle. A
ligation reaction using, for example, T4 DNA ligase is
performed under sufficiently dilute conditions so that
predominantly intramolecular ligations occur circularizing
the cDNA, with a only a minimum of intermolecular, concatamer
35 forming ligations. Reaction conditions favoring
circularization v rsus c ncatamer formation are described in
Maniatis, 1982, Molecular Cloning A Lab ratory Manual, pp.

- 120 -

124-125, 286-288, Cold Spring Harbor, NY. A DNA
concentration of less than approximately 1 μg/ml has been
found adequate to favor circularization. Concatamers can be
separated from circularized single molecules by size
5 separation using gel electrophoresis, if necessary. Fig. 4B
illustrates the circularized cDNA. Blunt end ligation
occurred between ends 407 and 408.

Then the circularized, biotin labeled, cDNA is cut
with REs and ligated to adapters uniquely recognizing and
10 perhaps uniquely labeled for each particular RE cut. The
RE/ligase step is performed by procedures described in the
sections hereinabove, for example in Sec. 5.2.2, so that RE
digestion and primer ligation proceed to completion with
minimal formation of concatamers and other unwanted ligation
15 products. Next, unwanted singly cut ends are removed by
contacting the reaction products with streptavidin or avidin
magnetic beads, leaving only doubly cut fragments that have
RE-specific recognition sequences ligated to each end. Fig.
4C illustrates these steps. Sequences 405 and 406 are cut by
20 RE_1 and RE_2, respectively, and adapters 421 and 422 specific
for cuts by RE_1 and RE_2, respectively are ligated onto the
overhangs. Thereby, fragment 409 is freed from the
circularized cDNA and adapters 421 and 422 are ligated to it.
The remaining segment of the circularized cDNA comprises
25 singly cut ends 423 and 424 with ligated adapters 421 and
422. Both singly cut ends are joined to the primer sequence
403 with attached biotin 404. Removal is accomplished by
contact with streptavidin or avidin 420 which is fixed to
substrate 425, perhaps comprising magnetic beads. Doubly cut
30 labeled fragment 409, now separated from the singly cut ends,
can be separated according to length and detected with
minimized background noise signals.

Thereby, signals from the labeled doubly cut ends
of interest can be directly detected with minimal
35 contamination from signals from unwanted labeled singly cut
ends. Importantly, the detect d signals more quantitatively
reflect th relative abundance of the source cDNA, and thus

- 121 -

gene expression levels. Optionally, if th signal l vels are too low for dir ct det ction, the r action products can be subjected to just the minimum number of cycles, for example according to the methods of Sec. 5.2.2, to detect the gene or

5 sequence of interest. For example, the number of cycles can be as small as four to eight without any concern of background contamination or noise. Thus, in this embodiment, amplification is not needed to suppress signals from singly cut ends, and preferred more quantitative response signal

10 intensities result.

Another QEA™ embodiment amplifies the cDNA sample prior to the RE/ligase reactions, removes unwanted fragments with a removal means, and then separates and detects the reaction products. Alternately, further amplification of the

15 fragments of interest can be performed after the RE/ligase step.

In this embodiment, first, double stranded cDNA, perhaps prepared from a tissue sample according to Sec. 6.3.1, is PCR amplified using primers a conjugated capture

20 moiety, preferably biotin. Any suitable primers known in the art, all biotin-labeled, can be used. For example, a set of arbitrary primers with no net sequence preference can be used. For a further example, where the cDNA is synthesized according to the protocol of Sec. 6.3.3, the method of step 6

25 of that protocol can be used, except that both the MA24 and MB24 have a conjugated biotin. The resulting cDNA with biotin linked to both ends is then cut with one or more REs and ligated to adapters corresponding to the REs used. The adapter primers can be optionally labeled but cannot have a

30 conjugated biotin. The RE/ligase reaction is preferably performed according to the protocols of Sec. 5.2.2 in order that the RE digestion and adapter ligation proceed to completion with minimum formation of concatamers and other unwanted ligation products. The reaction products comprise

35 fragments of interest that are doubly cut by REs and without any conjugated biotin, and unwanted fragments with a biotin conjugat d to one end that are singly cut and derive fr m the

ends of cDNAs. N xt, the unwanted singly cut fragments are
removed by contacting the reaction products with streptavidin
beads. Optionally, the purified fragments of interest can be
blunt-ended and subject to further PCR amplification for a
5 minimum number of cycles to observe the signals of interest.
Finally, the products are then analyzed, also as in the prior
sections, by separation according to length and by detection
of the DNA and of the optionally labeled adapter primers,
which indicate the RE cutting each fragment.
10   Other direct removal means may alternatively be
used in this invention. Such removal means include but are
not limited to digestion by single strand specific nucleases
or passage though a single strand specific chromatographic
column, for example, containing hydroxyapatite.
15   It will be apparent to those of skill in the art,
that these alternative protocols using cDNAs with a
conjugated capture moiety can combined with the other QEA™
embodiments in various manners. This invention encompasses
all such insubstantially different variations.
20

5.3. PCR EMBODIMENT OF QEA™

An alternative implementation of QEA™ methods not
using REs is based on PCR, or alternative amplification
means, to select and amplify cDNA fragments between chosen
25 target subsequences recognized by amplification primers.
See, generally, Innis et al., 1989, PCR Protocols A Guide to
Methods and Applications, Academic Press, New York, and Innis
et al., 1995, PCR Strategies, Academic Press, New York.
        Typically target subsequences between four and
30 eight base pairs long chosen by the methods previously
described are preferred because of their greater probability
of occurrence, and hence information content, as compared to
longer subsequences. However, DNA oligomers this short may
not hybridize reliably and reproducibly to their
35 complementary subsequences to be effectively used as PCR
primers. Hybridizati n reliability d pends strongly on
several variables, including primer composition and length,

stringency condition such as annealing temperature and salt
concentration, and cDNA mixtur c mplexity. For the hash
code to be effective for gene calling, it is highly preferred
that subsequence recognition be as specific and reproducible
5 as possible·so that well resolved bands representative only
cf the underlying sample sequence are produced. Thus,
instead of directly using single short oligonucleotides
complementary to the selected, target subsequences as
primers, it is preferable to use carefully designed primers.

10   The RE embodiments of QEA™ have been verified to
produce reproducible signal patterns over a 103 range on
input DNA concentrations. The PCR embodiment is less
preferred because the input DNA concentration, as well as the
initial hybridization temperature, must be closely to yield
15 reproducible results.

            The preferred primers are constructed according to
the mcdel in Fig. 5. Primer 501 is constructed of three
ccmponents, which, listed 5' to 3', are 504, 50ე, and 502.
Component 503, described infra, is optional. Component 502
20 is a sequence which is complementary to the subsequence which
primer 501 is designed to recognize. Component 502 is
typically 4-8 bp long. Component 504 is a 10-20 bp sequence
chosen so the final primer does not hybridize with any native
sequence in the cDNA sample to be analyzed; that is, primer
25 501 dces not anneal with any sequence known to be present in
the sample to be analyzed. The sequence of component 504 is
also chosen so that the final primer has a melting point
above 50°C, and preferably above 68°C. The method for
controlling melting temperature selecting average primer
30 composition and primer length is described above.

            Use of primer 501 in the PCR embodiment involves a
first annealing step, which allows the 3' end component 502
to anneal to its target subsequence in the presence of end
component 504, which may not hybridize. Preferably, this
35 annealing step is at a temperature between 36 and 44°C that
is empirically determined t maximiz r producibility of th
resulting signal pattern. The DNA concentration is

approximately 10 ng/50 ml and is similarly determined to
maximize r producibility. Other PCR conditions are standard
and are described in Sec. 6.6. Once annealed, the 3' end
serves as the primer elongation point for the subsequent
5 first elongation step. The first elongation step is
preferably at 72°C for 1 minute.

If stringency conditions are such that exact
complementarity is not required for hybridization, false
positive signals can be generated, that is signals resulting
10 from inexact recognition of the target subsequence. The
generation of these false positive bands can be accounted for
in the experimental analysis methods in order that DNA sample
sequences can still be recognized, but, perhaps, with some
increased recognition ambiguity that may need resolution.
15 These bands are accounted for by allowing inexact
hybridization matches of the target subsequence, the degree
of inexactness depending on the stringency of the
hybridization conditions. In this case the signals generated
contain only a fuzzy representation of the actual subsequence
20 in the sample, the degree of fuzziness being a function of
subsequence length and the stringency condition, that is
binding free energy, and the temperature of the
hybridization. Given the free energy and temperature, the
various possible actual subsequences can be approximately
25 determined by well known thermodynamic equilibrium
calculations.

Subsequent PCR cycles then use high temperature,
high stringency annealing steps. The high stringency
annealing steps ensure exact hybridization of the entire
30 primer. No further false positive bands are generated.
Preferably, these PCR cycles alternate between a 65°C
annealing step and 95°C melting step, each for 1 minute.

Optional component 503 can be used to improve the
specificity of the first low stringency annealing step and
35 thereby minimize false positive bands generated then.
Compon nt 503 can be $-(N)_j-$, where N is any nucleotide and j
is typically betw n 2 and 4, preferably 2. Use of all

- 125 -

possible c mponents 503 results in a degenerate set of
primers, 16 primers if j=2, which have a 3' end subsequence
effectively j bases longer than the target subsequence.
These longer complementary end sequences have improved
5 hybridization specificity. Alternately, component 503 can be
-(U)$_j$-, where N is a "universal" nucleotide and j is typically
between 2 and 4, preferably 3 or 4. A universal nucleotide,
such as inosine, is capable of forming base pairs with any
other naturally occurring nucleotide. In this alternative,
10 single primer 501 has a 3' end subsequence effectively j
bases longer than the target, and thus also has improved
hybridization specificity.

        A less preferred primer design comprises sets of
degenerate oligonucleotides of sufficient length to achieve
15 specific and reproducible hybridization, where each member of
a set includes a shared subsequence complementary to one
selected, target sequence. For example, if a subsequence to
be recognized is GATT, the set of primers used may be all
sequences of the form NNAATCNN, where N is any nucleotide.
20 Also sets of degenerate primers permit the recognition of
discontinuous subsequences. For example, GA--TT may be
recognized by all sequences of the form NAANNTCNN.
Alternately, a universal nucleotide can be used in place of
the degenerate nucleotides represented by 'N'.
25   Each primer or primer set used in a single reaction
is preferably distinctively labeled for detection. In the
preferred embodiment using electrophoretic fragment
separation, labeling is by fluorochromes that can be
simultaneously distinguished with optical detection means.
30   An exemplary experimental protocol is summarized
here, with details presented in Sec. 6.6. Total cellular
mRNA or purified sub-pools of cellular mRNA are used for cDNA
synthesis. First strand cDNA synthesis is performed
according to Sec. 6.3 using, for example, an oligo(dT) primer
35 or alternatively phasing primers. Alternatively, cDNA
sampl s can be prepared from any source or be directly
obtained.

Next, using a first strand cDNA sample, the primers
of the selected primer sets are used in a conventional PCR
amplification protocol. A high molar excess of primers is
preferably used to ensure only fragments between primer sites
5 that are adjacent on a target cDNA sequence or gene are
amplified. With a high molar excess of primers binding to
all available primer binding sites, no amplified fragment
should include internally any primer recognition site. As
many primers can be used in one reaction as can be labeled
10 for concurrent separation and detection and which generate an
adequately resolved length distribution, as in the RE
embodiments. For example, if fluorochrome labeling is used,
each pair of fluorochromes preferably is distinguishable in
one band and separate pairs preferably are distinguishable in
15 separate bands. After amplification, the fragments are
separated, re-suspended for gel electrophoresis,
electrophoretically separated, and optically detected.
Thereby the length distribution of fragments having
particular pairs of target subsequences at their ends is
20 ascertained.

Preferred protocols for the specific PCR
embodiments are described in detail in Sec. 6.6.

5.4. QEA™ ANALYSIS AND DESIGN METHODS

25   This inventions provides two groups of methods for
the Quantitative Expression Analysis embodiment of this
invention: first, methods for QEA™ experimental design; and
second, methods for QEA™ experimental analysis. Although,
logically, design precedes analysis, the methods of
30 experimental design depend on basic methods described herein
as part of experimental analysis. Consequently, experimental
analysis methods are described first.

In the following, descriptions are often cast in
terms of the preferred QEA™ embodiment, in which REs are used
35 to recognize target subsequences. However, such description
is not limiting, as all the methods to be described are
equally adaptable t all QEA™ mbodiments, including those in

which targ t subsequences are recogniz d by nucleic acid, or
nucl ic acid mimic, and probes which recognize target
subsequences by hybridization.

　　　　Further, the following descriptions are directed to
5 the currently preferred embodiments of these methods.
However, it will be readily apparent to those skilled in the
computer and simulation arts that many other embodiments of
these methods are substantially equivalent to those described
and can be used to achieve substantially the same results.
10 This invention comprises such alternative implementations as
well as its currently preferred implementation.

5.4.1.　OEA™ EXPERIMENTAL ANALYSIS METHODS

　　　　The analysis methods comprise, first, selecting a
15 database of DNA sequences representative of the DNA sample to
be analyzed, second, using this database and a description of
the experiment to derive the pattern of simulated signals,
contained in a database of simulated signals, which will be
produced by DNA fragments generated in the experiment, and
20 third, for any particular detected signal, using the pattern
or database of simulated signals to predict the sequences in
the original sample likely to cause this signal. Further
analysis methods present an easy to use user interface and
permit determination of the sequences actually causing a
25 signal in cases where the signal may arise from multiple
sequences, and perform statistical correlations to quickly
determine signals of interest in multiple samples.

　　　　The first analysis method is selecting a database
of DNA sequences representative of the sample to be analyzed.
30 In the preferred use of this invention, the DNA sequences to
be analyzed will be derived from a tissue sample, typically a
human sample examined for diagnostic or research purposes.
In this use, database selection begins with one or more
publicly available databases which comprehensively record all
35 observed DNA sequences. Such databases are GenBank from the
National Cent r for Biot chnology Inf rmati n (B thesda, MD),
the EMBL Data Library at the European Bioinformatics

Institute (Hinxton Hall, UK) and databas s from the National

Center for Genome Research (Santa Fe, NM). However, as any

sample of a plurality of DNA sequences of any provenance can

be analyzed by the methods of this invention, any database

5 containing entries for the sequences likely to be present in

such a sample to be analyzed is usable in the further steps

of the computer methods.

Fig. 6A illustrates the preferred database

selection method starting from a comprehensive tissue derived

10 database. Database 1001 is the comprehensive input database,

having the exemplary flat-file or relational structure 1010

shown in Fig. 6B, with one row, or record, 1014 for each

entered DNA sequence. Column, or field, 1011 is the

accession number field, which uniquely identifies each

15 sequence in database 1001. Most such databases contain

redundant entries, that is multiple sequence records are

present that are derived from one biological sequence.

Column 1013 is the actual nucleotide sequence of the entry.

The plurality of columns, or fields, represented by 1012

20 contain other data identifying this entry including, for

example whether this is a cDNA or gDNA sequence, if cDNA,

whether this is a full length coding sequence or a fragment,

the species origin of the sequence or its product, the name

of the gene containing the sequence, if known, etc. Although

25 shown as one file, DNA sequence databases often exits in

divisions and selection from all relevant divisions is

contemplated by this invention. For example, GenBank has 15

different divisions, of which the EST division and the

separate database, dbEST, that contain expressed sequence

30 tags ("EST") are of particular interest, since they contain

expressed sequences.

From the comprehensive database, all records are

selected which meet criteria for representing particular

experiments on particular tissue types. This is accomplished

35 by conventional techniques of sequentially scanning all

records in the c mprehensiv database, selecting th se that

- 129 -

match the criteria, and storing the selected records in a
selected database.

The following are exemplary selection methods. To
analyze a genomic DNA sample, database 1001 is scanned

5 against criteria 1002 for human gDNA to create selected
database 1003. To analyze expressed genes (cDNA sequences),
several selection alternatives are available. First, a
genomic sequence can be scanned in order to predict which
subsequences (exons) will be expressed. Thus selected

10 database 1005 is created by making selections according to
expression predictions 1004. Second, observed expressed
sequences, such as cDNA sequences, coding domain sequences
("CDS"), and ESTs, can be selected 1006 to create selected
database 1007 of expressed sequences. Additionally,

15 predicted and observed expressed sequences can be combined
into another, perhaps more comprehensive, selected database
of expressed sequences. Third, expressed sequences
determined by either of the prior methods may be further
selected by any available indication of interest 1008 in the

20 database records to create more targeted selected database
1009. Without limitation, selected databases can be composed
of sequences that can be selected according to any available
relevant field, indication, or combination present in
sequence databases.

25 The second analysis method uses the previously
selected database of sequences likely to be present in a
sample and a description of an intended experiment to derive
a pattern of the signals which will be produced by DNA
fragments generated in the experiment. This pattern can be

30 stored in a computer implementation in any convenient manner.
In the following, without limitation, it is described as
being stored as a table of information. This table may be
stored as individual records or by using a database system,
such as any conventionally available relational database.

35 Alternatively, the pattern may simply be stored as the image
of th in-memory structures which repr s nt the pattern.

A QEA™ experiment comprises several independent recognition reactions applied to th DNA sample sequ nces, where in each of the reactions labeled DNA fragments are produced from sample sequences, the fragments lying between 5 certain target subsequences in a sample sequence. The target subsequences can be recognized and the fragments generated by the preferred RE embodiments of QEA™ methods or by the PCR embodiment of QEA™. The following description is focused on the RE embodiments.

10   Fig. 7 illustrates an exemplary description 1100 of a preferred QEA™ embodiment. Field 1101 contains a description of the tissue sample which is the source of the DNA sample. For example, one experiment could analyze a normal prostrate sample; a second otherwise identical 15 experiment could analyze a prostrate sample with premalignant changes; and a third experiment could analyze a cancerous prostate sample. Differences in gene expression between these samples then relate to the progress of the cancer disease state. Such samples could be drawn from any other 20 human cancer or malignancy.

Major rows 1102, 1105, and 1109 describe the separate individual recognition reactions to which the DNA from tissue sample 1101 is subjected. Any number of reactions may be assembled into an experiment, from as few as 25 one to as many as there are pairs of available recognition means to recognize subsequences. Fig. 7 illustrates 15 reactions. For example, reaction 1 specified by major row 1102 generates fragments between target subsequences which are the recognition sites of restriction endonucleases 1 and 30 2 described in minor rows 1103 and 1104. Further, the RE1 cut end is recognized by a labeling moiety labeled with LABEL1, and the RE2 end is recognized by LABEL2. Similarly, reaction 15, 1109, utilizes restriction endonucleases 36 and 37 labeled with labels 3 and 4, minor rows 1110 and 1111, 35 respectively.

Maj r row 1105 d scribes a variant QEA™ r action using three REs and a s parate prob . As described, many REs

can be used in a single recognition r action as long as a
useful fragm nt distributi n results. Too many REs results
in a compressed length distribution. Further, probes for
target subsequences that are not intended to be labeled
5 fragment ends, but rather occur within a fragment, can be
used. For example, a labeled probe added after QEA™ PCR
amplification step (if present in a given embodiment), a post
PCR probe, can recognize subsequences internal to a fragment
and thereby provide an additional signal which can be used to
10 discriminate between two sample sequences which produce
fragments of the same length and end sequence which otherwise
have differing internal sequences. For another example, a
probe added before QEA™ PCR step and which cannot be extended
by DNA polymerase will prevent PCR amplification of those
15 fragment containing the probe's target subsequences. If PCR
amplification is necessary to generate detectable signals (in
a given embodiment), such a probe will prevent the detection
of such a fragment. The absence of a fragment may make a
previously ambiguous detected band now unambiguous. Such PCR
20 disruption probes can be PNA oligomers or degenerate sets of
DNA oligomers, modified to prevent polymerase extension
(e.g., by incorporation of a dideoxynucleotide at the 3'
end).

        Where alternative phasing PCR primers are used,
25 their extra recognition subsequences and labeling are
described in rows dependent to the RE/ligase reaction whose
products they are used to amplify.

        Next Fig. 8A illustrates, in general, that from the
database selected to best represent the likely DNA sequences
30 in the sample analyzed, 1201, and the description of QEA™
experiment, 1202, the simulation methods, 1203, determine a
pattern of simulated signals stored in a simulated database,
1204, that represents the results of QEA™ experiments. The
experimental simulation generates the same fragment lengths
35 and end subsequences from the input database that will be
generated in an actual xperiment p rformed on the same
sample f DNA s quences.

        Alternately, th simulat d pattern or database may
not be needed, in which case the DNA database is searched
sequence by sequence, mock digestions are performed and
compared against the input signals. A simulated database is
5 preferable if several signals need to be searched or if the
same QEA™ experiment is run several times. Conversely, the
simulated database can be dispensed with when few signals
from a few experiments need to searched. A quantitative
statement of when the simulated database is more efficient
10 depends upon an analysis of the costs of the various
operations and the size of DNA database, and can be performed
as is well known in the computer arts. Without limitation,
in the following the simulated database is described

        Fig. 8B illustrates an exemplary structure for the
15 simulated database. Here, the simulated results of all the
individual recognition reactions defined for the experiment
are gathered into rectangular table 1210. The invention is
equally adaptable to other database structures containing
equivalent information; such an equivalent structure would be
20 one, for example, where each reaction was placed in a
separate table. The rows of table 1210 are indexed by the
lengths of possible fragments. For example, row 1211
contains fragments of length 52. The columns of table 1210
are indexed by the possible end subsequences and probe hits,
25 if any, in a particular experimental reaction. For example,
columns 1212, 1213, and 1214 contain all fragments generated
in reaction 1, R1, which have both end subsequences
recognized by RE1, one end subsequence recognized by RE1 and
the other by RE2, and both end subsequences recognized by
30 RE2, respectively. Other columns relate to other reactions
of the experiment. Finally, the entries in table 1210
contain lists of the accession numbers of sequences in the
database that give rise to a fragment with particular length
and end subsequences. For example, entry 1215 indicates that
35 only accession number A01 generates a fragment of length 52
with both end subsequences rec gnized by RE1 in R1.
Similarly, entry 1216 indicat s that accession numbers A01

and S003 generat a fragment of l ngth 151 with both end
subsequences r cogniz d by RE3 in reaction 2.

In alternative embodiments, the contents of the
table can be supplemented with various information. In one
5 aspect, this information can aid in the interpretation of
results produced by the separation and detection means used.
For example, if separation is by electrophoresis, then the
detected electrophoretic DNA length can be corrected to
obtain the true physical DNA length. Such corrections are
10 well known in the electrophoretic arts and depend on such
factors. as average base composition and fluorochrome labels.
One commercially available package for making these
corrections is Gene Scan Software from Applied Biosystems,
Inc. (Foster City, CA). In this case, each table entry for a
15 fragment can contain additionally average base composition,
perhaps expressed as percent G+C content, and the
experimental definition can include primer average base
composition and fluorochrome label used. For a further
example, if separation is by mass spectroscopy or similar
20 method, the additional information can be the molecular
weight of each fragment and perhaps a typically fragmentation
pattern. Use of other separation and detection means can
suggest the use of other appropriate supplemental data.

Where alternative phasing primers, the SEQ-QEA™
25 embodiment, or other means generating effective targer.
subsequences are used, supplemental columns are used with RE
pair in order to further identify such effective target
subsequence.

Before describing how this simulated database is
30 generated, it is useful first to describe how this database
is used to predict experimental results. Returning to Fig.
7, labels are used to detect binding reaction events by
subsequence recognition means to the target DNA, to allow
detection after separation of the fragments by length. In an
35 embodiment using fluorescent detection means, these labels
are flu rochromes covalently attached to the primer strands
of the adapt rs, as previously described, or to hybridization

probes, if any. Typically, all the fluorochrom labels used
in on reaction are simultaneously distinguishable s that
fragments with all possible combinations of target
subsequences can be fluorescently distinguished. For
5 example, fragments at entry 1217 in table 1210 (Fig. 8B)
occur at length 175 and present simultaneous fluorescent
signals LABEL1 and LABEL2 upon stimulation, since these are
the labels used with adapters which recognize ends cuts by
RE1 and RE2 respectively. For a further example, in reaction
10 2, major row 1105 of experimental definition 1100 (Fig. 7), a
fragment with ends cut by RE2 and RE3 and hybridizing with
probe P will present simultaneous signals LABEL2, LABEL3, and
LABEL4. Where effective target subsequences are constructed.
e.g. by SEQ-QEA™ or alternative phasing primers, this lookup
15 is appropriately modified.

Other labelings are within the scope of this
invention. For example, a certain group of target
subsequences can be identically labeled or not labeled at
all, in which case the corresponding group of fragments are
20 not distinguishable. In this case, if RE1 and RE3 end
subsequences were identically labeled in table 1210 (Fig.
8B), a fragment of length 151 may be generated by sequence
T162, A01, or S003, or any combination of these sequences.
In the extreme, if silver (Ag) staining of an electrophoresis
25 gel is used in an embodiment to detect separated fragments,
then all bands will be identically labeled and only band
lengths can be distinguished within one electrophoresis lane.

Thus the simulated database together with the
experimental definition can be used to predict experimental
30 results. If a signal is detected in a recognition reaction,
say Rn, whose end labelings are LABEL1 and LABEL2 and whose
representation of length is corrected to physical length in
base pairs of L, the length L row of the simulated database
is retrieved and it is scanned for Rn entries with the
35 detected subsequence labeling, by using the column headings
indicating observed subsequences and the exp rimental
definition indicating how each subsequence is labeled. If n

- 135 -

match is found, this fragment represents a new gene or
sequenc not present in the selected database. If a match is
found, then this fragment, in addition to possibly being a
new gene or sequence, can also have been generated by those
5 candidate sequences present in the table entry(ies) found.

    The simulated database lookup is described herein
as using the physical length of a detected fragment. In
cases where the separation and detection means returns an
approximation to the true physical fragment length, lookup is
10 augmented to account for such as approximation. For example,
electrophoresis, when used as the separation means, returns
the electrophoretic length, which depending on average base
composition and labeling moiety is typically within 10% of
the physical length. In this case database lookup can search
15 all relevant entries whose physical length is within 10% of
the reported electrophoretic length, perform corrections to
obtain electrophoretic length, and then check for a match
with the detected signal. Alternative lookup implementations
are apparent, one being to precompute the electrophoretic
20 length for all predicted fragments, construct an alternate
table index over the electrophoretic length, and then
directly lookup the electrophoretic length. Other separation
and detection means can require corresponding augmentations
to lookup to correct for their particular experimental biases
25 and inaccuracies. It is understood that where database
lookup is referred to subsequently, either simple physical
lookup or augmented lookup is meant as appropriate.

    If matched candidate database sequences are found,
then the selected database can be consulted to determine
30 other information concerning these sequences, for example,
gene name, tissue origin, chromosomal location, etc. If an
unpredicted fragment is found, this fragment can be
optionally retrieved from the length separation means, cloned
or sequenced, and used to search for homologues in a DNA
35 sequence database or to isolate or characterize the
previously unknown gene or sequence. In this manner this

invention can b us d to rapidly discover and identify new
genes.

The computer methods of this invention are also
adaptable to other formats of an experimental definition.
5 For example, the labeling of the target subsequence
recognition moieties can be stored in a table separate from
the table defining the experimental reactions.

Now turning to the methods by which the simulated
database is generated, Fig. 9 illustrates a basic method,
10 termed herein mock fragmentation, which takes one sequence
and the definition of one reaction of an experiment and
produces the predicted results of the reaction on that
sequence. Generation of the entire simulated database
requires repetitive execution of this basic method.

15   Turning first to a description of mock
fragmentation, the method commences at 1301 and at 1302 it
inputs the sequence to be fragmented and the definition of
the fragmentation reaction, in the following terms: the
target end subsequences RE1 ... REn. where n is typically 2
20 or 3, and the subsequences to be recognized by post PCR
probes, P1 ... Pn, where n is typically 0 or 1. Note that
PCR disruption probes act as unlabeled end subsequences and
are so treated for input to this method. The operation of
the metnod is illustrated by example in Fig. 10A-F for the
25 case RE1, RE2 and P1.

At step 1303, for each target end subsequence, the
method makes a "vector of ends", which has elements which are
pairs of nucleotide positions along the sequence, each pair
being labeled by the corresponding end subsequence. For
30 embodiments where end subsequences are recognized by
hybridizing oligonucleotides, the first member of each pair
is the beginning of a target end subsequence and the second
member is the end of a target end subsequence. For
embodiments where target end subsequences are recognized by
35 restriction endonucleases, the first member of each pair is
th beginning f the verhang region that corresp nds to the
RE recognition subsequence and the second member is the end

of that overhang r gion. It is preferred to use REs that
generate 4 bp overhangs. The actual target end subsequences
are the RE recognition sequences, which are preferably 4-8 bp
long.

5   This vector is generated by a string operation
which compares the target end subsequence in a 5' to 3'
direction against the input sequence and seeks string
matches, that is the nucleotides match exactly. Where
effective target subsequences are formed by using, e.g. SEQ-
10 QEA™ or alternative phasing primers, it is the effective
subsequences that are compared. This can be done by simply
comparing the end subsequence against the input sequence
starting at one end and proceeding along the sequence one
base at time. However, it is preferable to use a more
15 efficient string matching algorithm, such as the Knuth-
Morris-Pratt or the Boyer-Moore algorithms. These are
described with sample code in Sedgewick, 1990, Algorithms in
C, chap. 19, Addison-Wesley, Reading, MA.
           In QEA™ embodiments where target subsequence are
20 recognized with accuracy, such as the RE embodiments, the
comparison of target subsequence against input sequence
should be exact, that is the bases should match in a one-to-
one manner. In embodiments where target subsequences are
less accurately recognized, the string match should be done
25 in a less exact, or fuzzy, manner. For example, in the PCR
embodiments, a target subsequence of length T can
inaccurately recognize an input sequence, also of length T,
by matching only T-n bases exactly, where n is typically 1 or
2 and is adjustable depending on experimental conditions. In
30 this case the string operation, which generates the vector of
ends, should accept partial T-n matches as well as exact
matches. In this, the string operations generate the false
positive matches expected from the experiments and permit
these fragments to be identified. Ambiguity in the simulated
35 database, however, increases, since more fragments leads to a
greater chance of fragments f identical length and nd
labels.

Fig. 10A illustrates nd vectors 1401 and 1402, comprising three and two ends, respectively, generated by RE1 and RE2, which are for this example assumed to be REs with a 4 bp overhang. The first overhang in vector 1401 occurs
5 between nucleotide 10 and 14 in the input sequence.

Step 1304 of Fig. 9 merges all the end vectors for all the end subsequences and sorts the elements on the position of the end. Vector 1404 of Fig. 10B illustrates the result of this step for example end vectors 1401 and 1402.
10   Step 1305 of Fig. 9 then creates the fragments generated by the reaction by selecting the parts of the full input sequence that are delimited by adjacent ends in the merged and sorted end vector. Since the experimental conditions in conducting QEA™ should be selected such that
15 target end subsequence recognition is allowed to go to completion, all possible ends are recognized. For the restriction endonuclease embodiments, the cutting and ligase reactions should be conducted such that all possible RE cuts are made and to each cut end a labeled primer is ligated.
20 These conditions insure that no fragments contain internal unrecognized target end subsequences and that only adjacent ends in the merged and sorted vector define generated fragments.

Where additional information is needed for
25 simulated database entries to adapt to inaccuracies in particular separation and detection means, such information can be collected at this step. For example, in the case of electrophoretic separation, fragment sequence can be determined and percent G+C content computed and entered in
30 the database along with the fragment accession number.

For the PCR embodiments, the fragment length is the difference between the end position of the second end subsequence and the start position of the first end subsequence. For RE embodiments, the fragment length is the
35 difference between the start position of the second end subsequence and the start p siti n of the first end

subsequence plus twice the primer length (48 in the preferred primer embodiment).

Fig. 10C illustrates the exemplary fragments generated, each fragment being represented by a 4 member
5 tuple comprising: the two end subsequences, the length, and an indicator whether the probe binds to this fragment. In Fig. 10C the position of this indicator is indicated by a '*'. Fragment 1408 is defined by ends 1405 and 1406, and fragment 1409 by ends 1406 and 1407. There is no fragment
10 defined by ends 1405 and 1407 because the intermediate end subsequence is recognized and either fully cut in an RE embodiment or used as a fragment end priming position in a PCR embodiment. For simplicity, the fragment lengths are illustrated for the RE embodiment without the primer length
15 addition.

Step 1306 of Fig. 9 checks if a hybridization probe is involved in the experiment. If not, the method skips to step 1309. If so, step 1307 determines the sequence of the fragment defined in step 1305. Fig. 10D illustrates that the
20 fragment sequences for this example are the nucleotide sequences within the input sequence that are between the indicated nucleotide positions. For example, the first fragment sequence is the part of the input sequence between positions 10 and 62. Step 1308 then checks each probe
25 subsequence against each fragment sequence to determine whether there is any match (i.e., whether the probe has a sequence complementary enough to the fragment sequence sufficient for it to hybridize thereon). If a match is found, an indication is made in the fragment 4 member tuple.
30 This match is done by string searching in a similar manner to that described for generation of the end vectors.

Next at step 1309 of Fig. 9, all the fragment are sorted on length and assembled into a vector of sorted fragments, which is output from the mock fragmentation method
35 at step 1310. This vector contains the complete list of all fragments, with probe information, defined by their end

subsequences and lngths that the input reaction will
gen rate from the input sequence.

Fig. 10E illustrates the fragment vector of the
example sorted according to length. For illustrative
5 purposes, probe P1 was found to hybridize only to the third
fragment 1412, where a 'Y' is marked. 'N' is marked in all
the other fragments, indicating no probe binding.

The simulated database is generated by iteratively
applying the basic mock fragmentation method for each
10 sequence in the selected database and each reaction in the
experimental definition. Fig. 11 illustrates a simulated
database generation method. The method starts at 1501 and at
1502 inputs the selected representative database and the
experimental definition with, in particular, the list of
15 reactions and their related subsequences. Step 1503
initializes the digest database table so that lists of
accession numbers may be inserted for all possible
combinations of fragment length and target end subsequences.
Step 1504, a DO loop, causes the iterative execution of steps
20 1505, 1506, and 1507 for all sequences in the input selected
database.

Step 1505 takes the next sequence in the database,
as selected by the enclosing DO loop, and the next reaction
of the experiment and performs the mock fragmentation method
25 of Fig. 9, on these inputs. Step 1506 adds the sorted
fragment vector to the simulated database by taking each
fragment from the vector and adding the sequence accession
number to the list in the database entry indexed by the
fragment length and end subsequences and probe (if any).
30 Fig. 10F represents the simulated database entry list
additions that would result for the example mock
fragmentation reaction of Figs. 10A-E. For example,
accession number A01 is added to the accession number list in
the entry 1412 at length 151 and with both end subsequences
35 RE2.

Finally, step 1507 t sts whether th re is another
reaction in the input experiment that should be simulated

- 141 -

against this sequence. If s , st p 1505 is repeated with
this r acti n. If not, th DO loop is repeated to select
another database sequence. If all the database sequences
have been selected, the step 1508 outputs the simulated
5 database and the method ends at 1509.

5.4.2. QEA™ EXPERIMENTAL DESIGN METHODS

The goal of the experimental design methods is to
optimize each experiment in order to obtain the maximum
10 amount of quantitative information. An experiment is defined
by its component recognition reactions, which are in turn
defined by the target end subsequences recognized, probes
used, if any, and labels assigned. If alternative phasing
primers, SEQ-QEA™, or other similar means are used, effective
15 target subsequences are used. Any of several criteria can.be
used to ascertain the amount of information obtained, and any
of several algorithms can be used to perform the reaction
optimization.

A preferred criteria for ascertaining the amount of
20 information uses the concept of "good sequence." A good
sequence for an experiment is a sequence for which there is
at least one reaction in the experiment that produces a
unique signal from that sequence, that is, a fragment is
produced from that good sequence, by at least one recognition
25 reaction, that has a unique combination of length and
labeling. For example, returning to Fig. 8B, the sequence
with accession number A01 is a good sequence because reaction
1 produces signal 1215, with length 52 and with both target
end subsequences recognized by RE1, uniquely from sequence
30 A01. However, sequence S003 is not a good sequence because
there are no unique signals produced only from S003: reaction
R2 produces signal 1216 from both A01 and S003 and signal
1219 from both Q012 and S003. Using the amount of good
sequences as an information measure, the greater the number
35 of good sequences in an experiment the better is the
xperimental d sign. Id ally, all p ssibl s quenc s in a
sample would be g od sequences.

Further, a quantitative measure of the expression of a good s quenc can simply be det rmin d from the detected signal intensity of the fragment uniquely produced from the good sequence. Relative quantitative measures of the
5 expression of different good sequences can be obtained by comparing the relative intensities of the signal uniquely produced from the good sequences. An absolute quantitative measure of the expression of a good sequence can be obtained by including a concentration standard in the original sample.
10 Such a standard for a particular experiment can consist of several different good sequences known not to occur in the original sample and which are introduced at known concentrations. For example, exogenous good sequence 1 is added at a $1:10^3$ concentration in molar terms; exogenous good
15 sequence 2 at a $1:10^4$ in molar terms; etc. Then comparison of the relative intensity of the unique signal of a good sequence in the sample with the intensities of the unique signal of the standards allows determination of the molar concentrations of the sample sequence. For example, if the
20 good sequence has a unique signal intensity half way between the unique signal intensities of good sequences 1 and 2, then it is present at a concentration half way between the concentrations of good sequences 1 and 2.

Another preferred measure for ascertaining the
25 amount of information produced by an experiment is derived by limiting attention to a particular set of sequences of interest, for example a set of known oncogenes or a set of receptors known or expected to be present in a particular tissue sample. An experiment is designed according to this
30 measure to maximize the number of sequences of interest that are good sequences. Whether other sequences possibly present in the sample are good sequences is not considered. These other sequences are of interest only to the extent that the sequences of interest produce uniquely labeled fragments
35 without any contribution from these other sequences.

This inv ntion is adaptabl to other measures f r ascertaining inf rmation from an experiment. For example,

- 143 -

another m asure is to minimiz on average the number of
sequenc s contributing t each detected signal. A further
measure is, for example, to minimize for each possible
sequence the number of other sequences that occur in common
5 in the same signals. In that case each sequence is linked by
common occurrences in fragment labelings to a minimum number
of other sequences. This can simplify making unambiguous
signal peaks of interest (see infra).

       Having chosen an information measure, for example
10 the number of good sequences, for an experiment, the
optimization methods choose target subsequences, and possibly
probes, which optimize the chosen measure. One possible
optimization method is exhaustive search, in which all
subsequences in lengths less than approximately 10 are tested
15 in all combinations for that combination which is optimum.
This method requires considerable computing power, and the
upper bound is determined by the computational facilities
available and the average probability of occurrence of
subsequences of a given length. With adequate resources, it
20 is preferable to search all sequences down to a probability
of occurrence of about 0.005 to 0.01. Upper bounds may rang
from 8 to 11 or 12.

       A preferred optimization method is known as
simulated annealing. See Press et al., 1986, Numerical
25 Recipes - The Art of Scientific Computing, Sec. 10.9,
Cambridge University Press, Cambridge, U.K. Simulated
annealing attempts to find the minimum of an "energy"
function of the "state" of a system by generating small
changes in the state and accepting such changes according to
30 a probabilistic factor to create a "better" new state. While
the method progresses, a simulated "temperature", on which
the probabilistic factor depends and which limits acceptance
of new states of higher energy, is slowly lowered.

       In the application to the methods of this
35 invention, a "state", denoted by S, is the experimental
definiti n, that is the targ t end subsequences and
hybridizati n probes, if any, in each recognition reaction of

- 144 -

the experiment. The "energy", d noted E, is taken to be 1.0
divided by the information measure, so that when the energy
is minimized, the information is maximized. Alternatively,
the energy can be any monotonically decreasing function of
5 the information measure. The computation of the energy is
denoted by applying the function E( ) to a state.

The preferred method of generating a new
experiment, or state, from an existing experiment, or state,
is to make the following changes, also called moves to the
10 experimental definition: (1) randomly change a target end
subsequence in a randomly chosen recognition reaction; (2)
add a randomly chosen target end subsequence to a randomly
chosen reaction; (3) remove a randomly chosen target end
subsequence from a randomly chosen reaction with three or
15 more target subsequences; (4) add a new reaction with two
randomly chosen target end subsequences; and (5) remove a
randomly chosen reaction. If an RE embodiment of QEA™ is
being designed, all target end subsequences are limited to
available RE recognition sequences. If alternative phasing
20 primers, SEQ-QEA™, or other means are used to generate
effective target subsequences, all subsequences must be
chosen from among such effective target subsequences that can
be generated from available REs. To generate a new
experimental definition, one of these moves is randomly
25 selected and carried out on the existing experimental
definition. Alternatively, the various moves can be
unequally weighted. In particular, if the number of
reactions is to be fixed, moves (4) and (5) are skipped. The
invention is further adaptable to other moves for generating
30 new experiments. Preferable generation methods will generate
all possible experiments.

Several additional subsidiary choices are needed in
order to apply simulated annealing. The "Boltzman constant"
is taken to be 1.0, so that the energy equals the
35 temperature. The minimum of the energy and temperature,
denot d E_0 and T_0, r spectively, ar d fined by the maximum of
th informati n m asur . For xample, if the number of good

- 145 -

sequenc s of interest is G and is used as the information
measure, then E_0, which equals T_0, equals 1/G. An initial
temperature, denoted T_1, is preferably chosen to be 1. An
initial experimental definition, or state, is chosen, either
5 randomly or guided by prior knowledge of previous
experimental optimizations. Finally, two execution
parameters are chosen. These parameters define the
"annealing schedule", that is the manner in which the
temperature is decreased during the execution of the
10 simulated annealing method. They are the number of
iterations in an epoch, denoted by N, which is preferably
taken to be 100 and the temperature decay factor, denoted by
f, which is preferably taken to be 0.95. Both N and f may be
systematically varied case-by-case to achieve a better
15 optimization of the experiment definition with a lower energy
and a higher information measure.

        With choices for the information measure or energy
function, the moves for generating new experiments, an
initial state or experiment, and the execution parameters
20 made as above, the general application of simulated annealing
to optimize an experimental definition is illustrated in Fig.
13A. The information measure used in this description is the
number of good sequences of interest. Any information
measure, such as those previously described, may be used
25 alternately.

        The method begins at step 1701. At step 1702 the
temperature is set to the initial temperature; the state to
the initial state or experimental definition; and the energy
is set to the energy of the initial state. At step 1703 the
30 temperature and energy are checked to determine whether
either is less than or equal to the minima for the
information measure chosen, as the result of either a
fortuitous initial choice or subsequent computation steps.
If the energy is less than or equal to the minimum energy, no
35 further optimization is possible, and the final experimental
definiti n and its energy is utput. If the t mperature is
less than or equal to the minimum t mperature, th

optimization is stopp d. Then the inverse of th energy is
the number of go d sequences of int rest for this
experimental definition.

Step 1706 is a DO loop which executes an epoch, or
5 N iterations, of the simulated annealing algorithm. Each
iteration consists of steps 1707 through 1711. Step 1707
generates a new experimental definition, or state, S_{new},
according to the described generation moves. Step 1708
ascertains or determines the information content, or energy,
10 of S_{new}. Step 1709 tests the energy of the new state, and, if
it is lower than the energy of the current state, at step
1711, the new state and new energy are accepted and replace
the current state and current energy. If the energy of the
new state is higher than the energy of the current state,
15 step 1710 computes the following function.

$$EXP[-(E-E_{new})/T]$$

This function defines the probabilistic factor controlling
acceptance. If this function is less than a random chosen
20 number uniformly distributed between 0 and 1, then the new
state is accepted at step 1711. If not, then the newly
generated state is discarded. These steps are equivalent to
accepting a new state if the energy is not increased by an
amount greater than that determined by function (4) in
25 conjunction with the selection of a random number. Or in
other words, a new state is accepted if the new information
measure is not decreased by an amount greater than indirectly
determined by function (4).

Finally, after an epoch of the algorithm, at step
30 1712 the temperature is reduced by the multiplicative factor
f and the method loops back to the test at step 1703.

Using this algorithm, starting from an initial
experimental definition which has certain information
content, the algorithm produces a final experimental
35 definition with a higher information content, or lower
energy, by r p titively and rand mly altering the

- 147 -

experimental definiti n in order to search for a d finition
with a high r information content.

5
The computation of the energy of an experimental
definition, or state, in step 1708 is illustrated more detail
in Fig. 13B. This method starts at step 1720. Step 1721
inputs the current experimental definition. Step 1722
determines a complete digest database from this definition
and a particular selected database by the method of Fig. 11.
Step 1723 scans the entire digest database and counts the
10 number of good sequences of interest. If the total number of
good sequences is the measure used, the total number of good
sequences can be counted. Alternatively, other information
measures may be applied to the digest database. Step 1724
computes the energy as the inverse of the information
15 measure. Alternatively, another decreasing function of the
information content may be used as the energy. Step 1725
outputs the energy, and the method ends at step 1726.

5.4.3. QEA™ AMBIGUITY RESOLUTION

20
In one utilization of this invention two related
tissue samples can be subject to the same experiment, perhaps
consisting of only one recognition reaction, and the outcomes
compared. The two tissue samples may be otherwise identical
except for one being normal and the other diseased, perhaps
25 by infection or a proliferative process, such as hyperplasia
or cancer. One or more signals may be detected in one sample
and not in the other sample. Such signals might represent
genetic aspects of the pathological process in one tissue.
These signals are of particular interest.
30
The candidate sequences that can produce a signal
of interest are determined, as previously described, by look-
up in the digest database. The signal may be produced by
only one sequence, in which case it is unambiguously
identified. However, even if the experiment has been
35 optimized, the signal may be ambiguous in that it may be
produced by sev ral candidate sequences fr m the select d

databas . A signal of inter st may be made unambiguous in
several manners which are described herein.

In a first manner of making unambiguous assume the
signal of interest is produced by several candidate sequences
5 all of which are good sequences for the particular
experiment. Then which sequences are present in the signal
of interest can be ascertained by determining the
quantitative presence of the good sequences from their unique
signals. For example, referring to Fig. 8B, if the signal
10 1217 of length 175 with the labeling 1213 is of interest, the
sequences actually present in the signal can be determined
from the quantitative determination of the presence of
signals 1215 and 1218. Here, both the possible sequences
contributing to this signal are good sequences for this
15 experiment.

The first manner of making unambiguous can be
extended to the case where cne of the sequences possibly
contributing to a signal is not a good sequence. The
quantitative presence of all the possible good sequences can
20 be determined from the quantitative strength of their unique
signals. The presence of the remaining sequence which is not
a good sequences can be determined by subtracting from the
quantitative presence of the signal of interest the
quantitative presences of all the good sequences.
25   Further extensions of the first manner can be made
to cases where more than one of the possible sequences is not
a good sequences if the sequences which are not good appear
as contributors to further signals involving good sequences
in a manner which allows their quantitative presences to be
30 determined. For example, suppose signal 1219 is of interest,
where both possible sequences are not good sequences. The
quantitative presence of sequence Q012 can be determined from
signals 1220 and 1218 in the manner previously outlined. The
quantitative presence of sequence S003 can be determined from
35 signals 1216 and 1215. Thereby, the sequences contributing
to signal 1219 can be determin d. More complex combinati ns
can be similarly made unambiguous.

An alt rnativ extension of the first mann r of
making unambiguous is by designing a further experiment in
which the possible sequences contributing to a signal of
interest are good sequences even if they were not originally
5 so. Since there are approximately 50 suitable REs that can
be used in the RE embodiment of QEA™ (Section 6.2), there are
approximately 600 RE reaction pairs that can be performed,
assuming that half of the theoretical maximum of 1,250 (50 X
50 / 2 = 1,250) are not useable. Since most RE pairs produce
10 on the average of 2CO fragments and standard electrophoretic
techniques can resolve at least approximately 500 fragment
lengths per lane, the RE QEA™ embodiment has the potential of
generating over 100,000 signals (500 X 200 = 100,000). The
number of possible signals is further increased by the use of
15 reactions with three or more REs and by the use of labeled
probes. Further, since the average complex human tissue, for
example brain, is estimated to express no more than
approximately 25,000 genes, there is a 4 fold excess of
possible signals over the number of possible sequences in a
20 sample. Thus it is highly likely that for any signal of
interest, a further experiment can be designed and optimized
for which all possible candidates of the signal of interest
are good sequences. This design can be made by using the
prior optimization methods with an information measure the
25 sequences of interest in the signal of interest and starting
with an extensive initial experimental definition including
many additional reactions. In that manner, any signal of
interest can be made unambiguous.

A second manner of making unambiguous is by
30 automatically ranking the likelihood that the sequences
possibly present in a signal of interest are actually present
using information from the remainder of the experimental
reactions. Fig. 14 illustrates a preferred ranking method.
The method begins at step 1801 and at step 1802 inputs the
35 list of possible accession numbers in a signal of interest,
the experimental definiti n, and the actual exp rimental
results. DO-loop 1803 it rates nc f r each possible

acc ssion number. Step 1804 performs a simulated experim nt
by the method illustrated in Fig. 11 in which, however, only
the current accession number is acted on. The output is a
single sequence digest table, such as illustrated in Fig.
5 10F.

        Step 1805 determines a numerical score of ranking
the similarity of this digest table to the experimental
results. One possible scoring metric comprises scanning the
digest table for all fragment signals and adding 1 to the
10 score if such a signal appears also in the experimental
results and subtracting 1 from the score if such signal does
not appear in the experimental results. Alternate scoring
metrics are possible. For example, the subtraction of 1 may
be omitted.
15   Step 1806 sorts the numerical scores of the
likelihood that each possible accession number is actually
present in the sample. Step 1807 outputs the sorted list and
the method ends at step 1808.
        By this method likelihood estimates of the presence
20 of the various possible sequences in a signal of interest can
be determined.

                5.5. **COLONY CALLING**
        The colony calling embodiment recognizes and
25 classifies single, individual genes or DNA sequences by
determining the presence or absence of target subsequences.
No length information is determined. This embodiment is
directed to gene determination and classification of arrayed
samples or colonies, where each sample or colony contains or
30 expresses only one sequence or gene of interest and is
perhaps prepared from a tissue cDNA library. The presence or
absence of target subsequences in a colony is determined by
use of labeled hybridization recognition means, each of which
uniquely binds to one target subsequence. It is preferable
35 that this binding be highly specific and reproducible. Each
sample or colony, or an array f samples or colonies, is
assayed for the c ntained sequence by d termining which of

                        - 151 -

the set of probes recognizes and thus hybridiz s to target
subs quences in the sample(s) or col ny(ies). Each sample is
then characterized by a hash code, each bit of which
indicates which probes recognized subsequences, or hits, in a
5 particular sample. The sequence or gene in a sample is
determined from the hash code by computer implemented
methods.

The choice of the target subsequences is important.
For economical and rapid assay, the size of the set of
10 recognition means should be as small as possible, preferably
less than 50 elements and more preferably from 15 to 25
elements. Further, it is most preferable that all possible
sequences or genes are recognized and uniquely determined.
It is preferable that 90 to 95% of all possible sequences be
15 recognized, with each sequence being indistinguishable from,
or ambiguous with, at most one or two other sequences.
Therefore, each target subsequence preferably occurs
frequently enough to minimize the number of different
recognition means needed. For example, it is not practical
20 for this invention, directed to rapid gene classification, if
each probe recognized only a few genes and therefore
thousands of probes were needed. However, each target
subsequence preferably does not occur so frequently that its
presence conveys little information. For example, a probe
25 recognizing every gene conveys no information.

The optimal choice is for each target subsequence
to have a probability of occurrence in all the genes or
sequences that can appear in a sample or colony of
approximately 50%; a preferable choice is a probability of
30 occurrence between 10 and 50%. Typically for human cDNA
libraries, target subsequences of length 4 to 6 meet this
condition, as longer sequences occur too infrequently to make
useful hash codes. Additionally, the presence of one target
subsequence is preferably independent of the presence of any
35 other target subsequence in the same sequence or gene. These
two criteria nsure that a hash code for a sampl , consisting
of indications f which target subsequ nces are present, is

maximally likely to repres nt a unique g ne or DNA sequence
with minimum f wasted code words not specifying any gene.
Such a hash code is an efficient representation of sequences
or genes.

5   The maximal number of genes or sequences that can
be represented by a hash code is 2^n, where n is the number of
target subsequences. A simple test to determine whether the
target subsequences occur frequently enough in the expected
gene library is made by comparing the actual probabilities of
10 the two hash codes that have all target subsequences either
present or absent to the ideal probabilities of these codes.
If p is the probability that any target subsequence occurs in
a given sequence in the library, then probability that none
of the target subsequences occur in a random gene is $(1-p)^n$.
15 The closer the ratio $(1-p)^n/2^{-n}$ is to 1 the more efficient is
the code. Similarly, the closer $p^\gamma/2^{-n}$, the ratio of the
probabilities that all the target subsequences are present to
the ideal probability conveying maximum information, is to 1
the more efficient is the code. We see the optimal p is
20 close to 2^{-1}.

          The preferred method of selecting target
subsequences meeting the probability of occurrence and
independence criteria is to use a database containing
sequences generally expected to be present in the samples to
25 be analyzed, for example human GenBank sequences for human
tissue derived samples. From a sequence database, oligomer
frequency tables are compiled containing the frequencies of,
preferably, all 4 to 8-mers. From these tables, candidate
subsequences with the desired probability of occurrence are
30 selected. Each candidate target subsequence is then checked
for independent occurrence, by, for example, checking that
the conditional probability for a hit by any selected pair of
candidates is approximately the product of the probabilities
of the individual candidate hit probabilities. Candidate
35 target subsequences meeting both occurrence and independence
criteria are possible targ t subs quences. A suffici nt

numb r, typically 20, of any of these subsequences can be
sel ct d as target subsequences for a hash code.

Preferably, but optionally, the initially set of
target subsequences can be optimized, using information on
5 the actual occurrences of the initially selected target
subsequences in the sequence database, resulting in a set of
target subsequences selected which recognizes a maximum
number of genes with a minimum number of sequences and with a
minimum amount of recognition ambiguity. Alternatively, this
10 optimization can also be performed on a sub-set of the
database comprised of sequences or genes of particular
biological or medical interest, for example, the set of all
oncogenes or growth factors. In this manner, fewer target
subsequences can be chosen which distinguish more efficiently
15 among a set of sequences or genes of particular interest and
distinguish that set of genes from the sequences of the
remainder of the sample.

This combinatorial optimization problem is
computationally intensive to solve exactly. A number of
20 approximate techniques can be used to obtain efficient nearly
optimal solutions. The preferred but not limiting technique
is to use simulated annealing (Press et al., 1986, Numerical
Recipes - The Art of Scientific Computing, Sec. 10.9,
Cambridge University Press, Cambridge, U.K.). The
25 experimental design and optimization are described in detail
in the following section.

Example 6.6 illustrates the results of the
simulated annealing optimization method. Simulated annealing
generally produces a choice of subsequences that achieve the
30 same resolution while using approximately 20% fewer total
sequences than a selection guided only by the probability
principles previously described. This level of optimization
is likely to improve with larger and less redundant databases
that represent longer genes.

35 An alternative to using single target subsequences
is to use sets of target subsequences, r cognized by s ts of
identically label d hybridization probes, to generate one

presence or abs nce indication for the hash code. In this
alt rnative, sets of longer target subsequences would be
chosen such that the presence of any target subsequence in
the set is a presence indication. Absence means no element
5 of the set is present. If the sets are chosen so that their
probability of presence in a single sequence is near 50%,
preferably from 10 to 50%, and the presence or absence of one
set is independent of the presence or absence of any other
set, such sets can be used to construct codes equally well as
10 single subsequences. A resulting code will be efficient and
can be further optimized by simulated annealing, as for
single target subsequence codes. Target sets of longer
subsequences are preferable where experimental recognition cf
shorter subsequences is less specific and reproducible, as
15 for example is true where short DNA oligomers are used as
hybridization probes for recognition. As a further.
alternative, a code can consist of presence or absence
indications of mixed target sets of subsequences and single
target subsequences.
20   Probes for a target subsequence are preferably PNA
oligomers, or less preferably DNA oligomers, which hybridize
to the subsequence of interest. Use of sets of degenerate
DNA oligomers to more specifically and reliably hybridize to
short DNA subsequences has been described in relation to the
25 PCR implementation of QEA™ methods. The use of PNAs is
preferred in the colony calling embodiment since PNA
oligomers, due to their more favorable hybridization
energetics, more specifically and reliably hybridize to
shorter complementary DNA subsequences than do DNA oligomers.
30 Reliable hybridization occurs for PNA 6 to 8-mers and longer.
Probing shorter subsequences preferably uses fully degenerate
sets of PNA oligomers, as is the case for DNA oligomers.
        PNAs are even more preferable when, in the
alternative, the hash code comprises presence or absence
35 indication of target sets of longer subsequences. In this
case, many m r DNA probes are gen rally required than PNA
prob s. As PNA 6 to 8-m rs reliably hybridiz , target sets

can consist of subs quences of length 6 to 8. Since DNA oligom rs of this l ngth may not reliably hybridize, each subsequence in the set must in turn be represented by a further degenerate set of DNA oligomers, requiring thereby a
5 set of sets.

The experimental method of colony calling comprises three principal steps: first, arraying cDNA libraries on filters or other suitable substrates; second, PNA hybridization and detection, alternatively DNA hybridization
10 can be used; and third, interpreting the resulting hash code to determine the sequence in the sample.

The first step, which can be omitted if arrayed cDNA libraries are already available, is constructing and arraying cDNA libraries. Any methods known in the art may be
15 used. For example, cDNA libraries from normal or diseased tissues can be constructed according to Example 6.3. Alternatively, the human cDNA libraries constructed by M.B. Soares and colleagues are available as high density arrays on filters and can be used for the practice of this method. See
20 Scares et al., 1994, *Proc. Natl. Acad. Sci. USA* 91:9228-32. The ability to spot up to thousands of cDNA clones or colonies on filters suitable for hybridization is an established technology. This service is now provided by several companies, including the preferred supplier Research
25 Genetics (Huntsville, AL). The protocol of Example 6.7 can be used to generate these arrays from cDNA libraries.

The second step is probe (*e.g.*, PNA) hybridization and detection. Fluorescently labeled PNA oligomers are available from PerSeptive Biosystems (Bedford, MA) or can be
30 synthesized. PNAs are designed to be complementary to the chosen target subsequences and to have a maximum number of distinguishable labels for simultaneous hybridization with multiple oligomers. PNA hybridization is performed according to standard protocols developed by the manufacturer and
35 detailed in Example 6.7. Detection of the PNA signals uses optical spectrographic means to distinguish fluorochrome missions similar to those us d in DNA analysis instrum nts,

but appr priately modified to recognize spots on filters as
opposed t linearly arrayed bands.

The third step, interpretation of the hash code, is
done by the computer implemented method described in the
5 following section.

In an alternative embodiment, the intensity of the
detected hybridization signal indicates the number of times
the probe binds to the sample sequence. In this manner the
number of recognized target subsequences present in the
10 sample can be determined. This information can be used to
more precisely classify of identify a sample.

5.6.   CC ANALYSIS AND DESIGN METHODS

The colony calling ("CC") computer implemented
15 methods are similar to QEA™ computer methods. As for QEA™,
the experimental analysis methods are described before the
experimental design methods.

5.6.1.   CC EXPERIMENTAL ANALYSIS METHODS

20   The analysis methods make use of a mock experiment
concept. First, a database is selected to represent possible
sequences in the sample by the same methods as described for
QEA™ analysis. These are illustrated and described with
reference to Fig. 6A. For CC, an experimental definition is
25 simply a list of N_p target subsequences, where N_p is
preferably between 16 and 20. Next, a mock experiment
generates one hash code for each sequence in the selected
database, each hash code being a string of N_p binary digits
wherein the n'th digit is a 1 (0) if the n'th target
30 subsequence does (does not) hybridize with the sequence. The
results of all the mock experiments determine the pattern of
hash codes expected. This pattern is output in a code table
of all possible hash codes in which, for each hash code,
there is a list of all accession numbers of sequences with
35 this code.

This method is illustrated in more detail in Fig.
15. The method starts at step 1901 and at step 1902 it

inputs a s lected database and on exp rimental definiti n
consisting of N_p target subs quences. Step 1903 initializes a
table which for each of the 2^{N_p} hash codes can contain a list
of possible accession numbers which have this hash code.

5 Step 1904 is a DO loop which iterates through all sequences
in the database. For a particular sequence, step 1905 checks
for each target subsequence whether that subsequence
hybridizes to the sequence. This is implemented by string
matching in a manner similar to step 1303 of Fig. 9. A

10 binary hash code is constructed from this hybridization
information, and step 1906 adds the accession number of the
sequence to the list of accession numbers associated with
this hash code in the code table. Step 1907 outputs the code
table and the method ends at step 1908.

15   Having built a pattern of simulated hash code in a
code table, analysis of an experiment requires only simple
table look-up. A colony is hybridized with each of the N_p
recognition means for the target subsequences. The results
of the hybridization are used to construct a resulting hash

20 code. This code table for this hash code entry then contains
a list of sequence accession numbers that are possible
candidates for the sample sequence. If the list contains
only cne element, then the sample has been uniquely
identified. If the list contains more than one element, the

25 identification is ambiguous. If the list is empty, the
sample is not in the selected database and may possibly be a
previously unknown sequence.

          Alternately, as for QEA™ experimental analysis, a
code table can be dispensed with if only a few hash codes

30 need to be looked up from only a few experiments. Then the
DNA database is scanned sequence by sequence for those
sequences generating the hash code of interest. If many hash
codes from many experiments need to be analyzed, a code table
is more efficient. The quantitative decision of when to

35 build a code table depends on the costs of the various
operations and the siz of DNA database, and can be p rform d

as is well known in the computer arts. Without limitation,
this d scription is built on the use f a code table.

For those embodiments where the recognition means
can each recognize a subset of target subsequences, code
5 table construction must be modified accordingly. Such
embodiments, for example, can involve DNA oligomer probes
which due to their length can hybridize with an intended
target subsequences and those subsequences which differ by 1
base pair from the intended target. In such embodiments,
10 step 1905 checks whether each member of such a set of target
subsequences is found in the sample sequence. If any member
is found in the sequence, then this information is used to
construct the hash code.

15   5.6.2. CC EXPERIMENTAL DESIGN METHODS

As for QEA™, the goal of CC experimental design is
to maximize the amount of information from a CC hybridization
experiment. This is also performed by defining an
information measure and choosing an optimization method which
20 maximizes this measure.

The preferred information measure is the number of
occupied hash codes. This is equivalent to minimizing the
number of accession numbers which can result in a given hash
code. In fact for N_p greater than about 17 to 18, that is for
25 2^{N_p} greater than the number of expressed human genes (about
100,000), maximizing the number of occupied hash codes can
result in each hash code representing a single sequence.
Such a unique code contains the maximum amount of
information. The invention is adaptable to other CC
30 information measures. For example, if only a subset of the
possible sequences are of interest, an appropriate measure
would be the number of such sequences which are uniquely
represented by a hash code. As for QEA™, these are sequences
of interest.
35   One optimization algorithm is exhaustive search.
In exhaustive search, all subsequences of l ngth less than
approximately 10 ar tri d in all c mbinations in order to

find the optimum combination producing th b st hash code
according to th chosen information measure. This method is
inefficient. The preferred algorithm for optimizing the
information from an experiment is simulated annealing. This
5 is performed by the method illustrated and described with
respect to Fig. 13A. For CC, the following preferred choices
are made.

The energy is taken to be 1.0 divided by the
information content; alternatively, any monotonically
10 decreasing function of the information content can be used.
The energy is determined by performing the mock experiment of
Fig. 15 using a particular experimental definition and then
applying the measure to the resulting code table. For
example, if the number of occupied hash codes is the
15 information measure, this number can be computed by simply
scanning the code table and counting the number of table
entries with non-empty accession number lists. The Boltzman
constant is again taken to be 1 so that the temperature
equals the energy. The initial temperature is preferably
20 1.0. The minimum energy and temperature, E_0 and T_0,
respectively, are determined by the information measure. For
example, with the prior choices for energy function and
information measure, E_0, which equals T_0, is 1.0 divided by
the number of sequences in the selected database.
25   The method of generating a new experimental
definition from an existing definition is to pick randomly
one target subsequence and to perform one of the following
moves: (1) randomly modifying one or more nucleotides; (2)
adding a random nucleotide; and (3) removing a random
30 nucleotide. A modification is discarded if it results in two
identical target subsequences. Further, it is desirable to
discard a modification if the resulting subsequence has an
extreme probability of binding to sequences in the database.
For example, if the modified subsequence binds with a
35 probability less than approximately 0.1 or more than
approximately 0.5 t sequ nc s in the selected database, it
should be discarded. To generate a new experiment, one of

- 160 -

th se moves is rand mly selected and carried out on the
existing experimental definition. Alternatively, the various
moves can be unequally weighted. The invention is further
adaptable to other methods of generating new experiments.
5 Preferably, generation methods used will randomly generate
all possible experiments. An initial experimental definition
can be picked by taking N_p randomly chosen subsequences or by
using subsequences from prior optimization.

      Finally, the two execution parameters defining the
10 "annealing schedule", that is the manner in which the
temperature is decreased during the execution of the
simulated annealing method, are defined and chosen as in
QEA™. The number of iterations in an epoch, denoted by N, is
preferably taken to be 100 and the temperature decay factor,
15 denoted by f, is preferably taken to be 0.95. Both N and f
may be systematically varied case-by-case to achieve a better
experimental definition with lower energy and a higher
information measure.

      With these choices the simulated annealing
20 optimization method of Fig. 13A can be performed to obtain an
optimized set of target subsequences. To determine an
optimum N_p, different initial N_p can be selected, the prior
design optimization performed, and the results compared. The
Np with the maximum information measure is optimum for the
25 selected database.

5.6.3.   CC QUANTITATIVE EMBODIMENT

      To make use of quantitative detection information
the pattern of simulated hash codes stored in the code table
30 is augmented with additional information. For each hash code
in the table and each sequence giving rise to that hash code,
this additional information comprises recording the number of
times each target subsequence is found in such a sequence.
These numbers are simply determined by scanning the entire
35 sequence and counting the number of occurrences of each
target subsequ nce.

An exemplary method to perform hash code look up in
this augment d tabl is t first find the sequences giving
rise to a particular hash code as a binary number, and second
to pick from these the most likely sequence as that sequence
5 having the most similar pattern of subsequence counts to the
detected quantitative hybridization signal. An exemplary
method to determine such similarity is to linearly normalize
the detected signal so that the smallest hybridization signal
is 1.0 and then to find the closest sequence by using a
10 Euclidean metric in an n-dimensional code space.

For CC experimental design, each pattern of
subsequence counts may alternatively be considered as a
distinct code entry for evaluation of an information measure.
This is instead of considering each hash code alone a
15 distinct entry.

5.7. APPARATUS FOR PERFORMING THE METHODS OF THE INVENTION

The apparatus of this invention includes means for
performing the recognition reactions of this invention in a
20 preferably automated fashion, for example by the protocols of
§ 6.4.3, and means for performing the computer implemented
experimental analysis and design methods of this invention.
Although the subsequent discussion is directed to embodiments
of apparatus for QEA™ embodiments of this invention, similar
25 apparatus is adaptable to the CC embodiments. Such adaption
includes using, in place of the corresponding components for
QEA™ embodiments, automatic laboratory instruments
appropriate for making and hybridizing arrays of clones and
for reading the results of the hybridizations, and using
30 programs implementing the computer analysis and design
methods for the CC embodiments described in Sec. 5.6.

Fig. 12A illustrates an exemplary apparatus for
QEA™ embodiments of this invention, and with the described
adaption, also for the CC embodiments of this invention.
35 Computer 1601 can be, alternatively, a UNIX based work
station type c mput r, an MS-DOS or Windows based personal
computer, a Macintosh personal computer, or anoth r

- 162 -

equival nt computer. In a preferred embodiment, comput r
1601 is a Pow rPC™ based Macintosh computer with software
systems capable of running both Macintosh and MS-DOS/Windows
programs.

5   Fig. 12B illustrates the general software structure
in RAM memory 1650 of computer 1601 in a preferred
embodiment. At the lowest software level is Macintosh
operating system 1655. This system contains features 1656
and 1657 for permitting execution of UNIX programs and MS-DOS
10 or Windows programs alongside Macintosh programs in computer
1601. At the next higher software level are the preferred
languages in which the computer methods of this invention are
implemented. LabView 1658, from National Instruments
(Dallas, TX), is preferred for implementing control routines
15 1661 for the laboratory instruments, exemplified by 1651 and
1652, which perform the recognition reactions and fragment
separation and detection. C or C++ languages 1659 are
preferred for implementing experimental routines 1662, which
are described in Sec. 5.4 and 5.6. Less preferred, but
20 useful for rapid prototyping, are various scripting languages
known in the art. PowerBuilder 1660, from Sybase (Denver,
CO), is preferred for implementing the user interfaces to the
computer implemented routines and methods. Finally, at the
highest software level are the programs implementing the
25 described computer methods. These programs are divided into
instrument control routines 1661 and experimental analysis
and design routines 1662. Control routines 1661 interact
with laboratory instruments, exemplified by 1651 and 1652,
which physically perform QEA™ and CC protocols. Experimental
30 routines 1662 interact with storage devices, exemplified by
devices 1654 and 1653, which store DNA sequence databases and
experimental results.

          Returning to Fig. 12A, although only one processor
is illustrated, alternatively, the computer methods and
35 instrument control interface can be performed on a
multiprocessor or n several s parate but linked pr c ssors,
such that instrument control m thods 1661, computational

experimental methods 1661, and the graphical interface
methods can be on different processors in any combination or
sub-combination.

5   Input/output devices include color display device
1620 controlled by a keyboard and standard mouse 1603 for
output display of instrument control information and
experimental results and input of user requests and commands.
Input and output data are preferably stored on disk devices
such as 1604, 1605, 1624, and 1625 connected to computer 1601
10 through links 1606. The data can be stored on any
combination of disk devices as is convenient. Thereby, links
1606 can be either local attachments, whereby all the disks
can be in the computer cabinet(s), LAN attachments, whereby
the data can be on other local server computers, or remote
15 links, whereby the data can be on distant servers.

             Instruments 1630 and 1631 exemplify laboratory
devices for performing, in a partly or wholly automatic
manner, QEA™ recognition reactions. These instruments can
be, for example, automatic thermal cyclers, laboratory
20 robots, and controllable separation and detection apparatus,
such as is found in the applicants' copending U.S. Patent
Application 08/438,231 filed May 9, 1995. Links 1632
exemplify control and data links between computer 1601 and
controlled devices 1631 and 1632. They can be special buses,
25 standard LANs, or any suitable link known in the art. These
links can alternatively be computer readable medium or even
manual input exchanged between the instruments and computer
1601. Outline arrows 1634 and 1635 exemplify the physical
flow of samples through the apparatus for performing
30 experiments 1607 and 1613. Sample flow can be either
automatic, manual, or any combination as appropriate. In
alternative embodiments there may be fewer or more laboratory
devices, as dictated by the current state of the laboratory
automation art.
35   On this complete apparatus, a QEA™ experiment is
designed, performed, and analyz d, pr ferably in a manner as
automatic as possible. First, a QEA™ xp riment is designed,

according to the methods specified in Sec. 5.4.2 as
implemented by xperimental routines 1662 on computer 1601.
Input to the design routines are databases of DNA sequences,
which are typically representative selected database 1605
5 obtained by selection from input comprehensive sequence
database 1604, as described in Sec. 5.4.1. Alternatively,
comprehensive DNA databases 1604 can be used as input.
Database 1604 can be local to or remote from computer 1601.
Database selection performed by processor 1601 executing the
10 described methods generates one or more representative
selected databases 1605. Output from the experimental design
methods are tables, exemplified by 1609 and 1615, which, for
a QEA™ RE embodiment, specify the recognition reaction and
the REs used for each recognition reaction.
15   Second, the apparatus performs the designed
experiment. Exemplary experiment 1607 is defined by tissue
sample 1608, which may be normal or diseased, experimental
definition 1609, and physical recognition reactions 1610 as
defined by 1609. Where instrument 1630 is a laboratory robot
20 for automating reaction, computer 1601 commands and controls
robot 1630 to perform reactions 1610 on cDNA samples prepared
from tissue 1608. Where instrument 1631 is a separation and
detection instrument, the results of these reactions are then
transferred, automatically or manually, to 1631 for
25 separation and detection. Computer 1601 commands and
controls performance of the separation and receives detection
information. The detection information is input to computer
1601 over links 1632 and is stored on storage device 1624,
along with the experimental design tables and information on
30 the tissue sample source for processing. Since this
experiment uses, for example, fluorescent labels, detection
results are stored as fluorescent traces 1611.
            Experiment 1613 is processed similarly along sample
pathway 1633, with robot 1630 performing recognition
35 reactions 1616 on cDNA from tissue 1608 as defined by
definiti n 1615, and device 1631 performing fragment
separati n and detection. Fragment d tection data is input

- 165 -

by comput r 1601 and stored on storag devic 1625. In this
case, for example, silver staining is us d, and detection
data is image 1617 of the stained bands.

During experimental performance, instrument control
5 routines 1661 provide the detailed control signals needed by
instruments 1630 and 1631. These routines also allow
operator monitoring and control by displaying the progress of
the experiment in process, instrument status, instrument
exceptions or malfunctions, and such other data that can be
10 of use to a laboratory operator.

Third, interactive experimental analysis is
performed using the database of simulated signals generated
by analysis and design routines 1662 as described in Sec.
5.4.2 and 5.4.3. Simulated database 1612 for experiment 1607
15 is generated by the analysis methods executing on processor
1601 using as input the appropriate selected database 1605
and experimental definition 1609, and is output in table
1612. Similarly table 1618 is the corresponding simulated
database of signals for experiment 1613, and is generated
20 from appropriate selected database 1605 and experimental
definition 1615. A signal is made unambiguous by
experimental routines 1662 that implement the methods
described in Sec. 5.4.3.

Display device 1602 presents an exemplary user
25 interface for the data generated by the methods of this
invention. This user interface is programmed preferably by
using the Powerbuilder display front end. At 1620 are
selection buttons which can be used to select the particular
experiment and the particular reaction of the experiment
30 whose results are to be displayed. Once the experiment is
selected, histological images of the tissue source of the
sample are presented for selection and display in window
1621. These images are typically observed, digitized, and
stored on computer 1601 as part of sample preparation. The
35 results of the selected reaction of the selected experiment
ar displayed in window 1622. Here, a fluoresc nt trace
output of a particular labeling is made available. Window

1622 is indexed by marks 1626 representing the possible locations of DNA fragments of successive int g r lengths.

Window 1623 displays contents from simulated database 1612. Using, for example, mouse 1603, a particular
5 fragment length index 1626 is selected. The processor then retrieves from the simulated database the list of accession numbers that could generate a peak of that length with the displayed end labeling. This window can also contain further information about these sequences, such as gene name,
10 bibliographic data, etc. This further information may be available in selected databases 1605 or may require queries to the complete sequence database 1604 based on the accession numbers. In this manner, a user can interactively inquire into the possible sequences causing particular results and
15 can then scan to other reactions of the experiment by using buttons 1620 to seek other evidence of the presence of these sequences.

It is apparent that this interactive interface has further alternative embodiments specialized for classes of
20 users of differing interests and goals. For a user interested in determining tissue gene expression, in one alternative, a particular accession number is selected from window 1623 with mouse 1603, and processor 1601 scans the simulated database for all other fragment lengths and their
25 recognition reactions that could be produced by this accession number. In a further window, these lengths and reactions are displayed, and the user allowed to select further reactions for display in order to confirm or refute the presence of this accession number in the tissue sample.
30 If one of these other fragments are generated uniquely by this sequence (a "good sequence", see supra), that fragment can be highlighted as of particular interest. By displaying the results of the generating reaction of that unique fragment, a user can quickly and unambiguously determine
35 whether or not that particular accession number is actually present in the sample.

In anoth r interfac alt rnative, the system
displays two experiment side by side, displaying two
histological images 1621 and two experimental results 1622.
This allows the user to determine by inspection signals
5 present in one sample and not present in the other. If the
two samples were diseased and normal specimens of the same
tissue, such signals would be of considerable interest as
perhaps reflecting differences due to the pathological
process. Having a signal of interest, preferably repeatable
10 and reproducible, a user can then determine the likely
accession numbers causing it by invoking the previously.
described interface facilities. In a further elaboration of
this embodiment, system 1601 can aid the determination of
signals of interest by automating the visual comparison by
15 performing statistical analysis of signals from samples of
the same tissue in different states. First, signals
reproducibly present in tissue samples in the same state are
determined, and second, differences in these reproducible
signals across samples from the several states are compared.
20 Display 1602 then shows which reproducible signals vary
across the states, thereby guiding the user in the selection
of signals of interest.

The apparatus of this invention has been described
above in an embodiment adapted to a single site
25 implementation, where the various devices are substantially
local to computer 1601 of Fig. 12A, although the various
links shown could also represent remote attachments. An
alternative, explicitly distributed embodiment of this
apparatus is illustrated in Fig. 12C. Shown here are
30 laboratory instruments 1670, DNA sequence database systems
1684, and computer systems 1671 and 1673, all of which
cooperate to perform the methods of this invention as
described above.

These systems are interconnected by communication
35 medium 1674 and its local attachments 1675, 1676, and 1677 to
the various systems. This medium may be any dedicated or
shared or local or remote communication medium known in the

art. For example, it can be a "campus" LAN network extending
perhaps a few kilom t rs, a d dicated wide area communication
system, or a shared network, such as the Internet. The
system local attachments are adapted to the nature of medium
5 1674.   .

Laboratory instruments 1670 are commanded by
computer system 1671 to perform the automatable steps of the
recognition reactions, separation of the reaction results,
and detection and transmission of resulting signals through
10 link 1672. Link 1672 can be any local or remote link known
in the art that is adapted to instrument control, and may
even be routed through communication medium 1674..

DNA sequence database systems 1684 with various
sequence databases 1685 may be remote from the other systems,
15 for example, by being directly accessed at their sites of
crigin, such as Genbank at Bethesda, MD. Alternatively,
parts or all of these databases may be periodically
downloaded for local access by computer systems 1671 and 1672
onto such storage devices as discs cr CD-ROMs.
20   Computer system 1671, including computer 1681,
storage 1682, and display 1683, can perform various methods
of this invention. For example, it can perform solely the
control routine for control and monitoring of instrument
system 1670, whereby experimental design and analysis are
25 performed elsewhere, as at computer system 1673. In this
case, system 1671 it would typically be operated by
laboratory technicians. Alternatively, system 1671 can also
perform experimental designs, which meet the requirements of
remote users of sample analysis information. In another
30 embodiment, system 1671 can carry out all the computer
implemented methods of this invention, including final data
display, in which case it would be operated by the final
users of the analysis information.

Computer system 1673, including computer 1678,
35 storage 1679, and display 1680, can perform a corresponding
rang of functions. However, typically system 1673 is
r motely located and would be used by final users of the DNA

- 169 -

sample information. Such users can include clinicians seeking informati n to make a diagnosis, grade or stage a disease, or guide therapy. Other users can include pharmacologists seeking information useful for the design or
5 improvement of drugs. Finally, other users can include researchers seeking information useful to basic studies in cell biology, developmental biology, etc. It is also possible that a plurality of computer systems 1673 can be linked to laboratory system 1670 and control system 1671 in
10 order to provide for the analysis needs of a plurality of classes of users by designing and causing the performance of appropriate experiments.

It will be readily apparent to those of skill in the computer arts that alternative distributed embodiments of
15 the apparatus of this invention, along with alternative functional allocations of the computer implemented methods to the various distributed systems, are equally possible.

All the computer implemented methods of this invention can be recorded for storage and transport on any
20 computer readable memory devices known in the art. For example, these include, but are not limited to, semiconductor memories - such as ROMs, PROMs, EPROMs, EEPROMS, etc. of whatever technology or configuration - magnetic memories - such as tapes, cards, disks, etc of whatever density or size
25 - optical memories - such as optical read-only memories, CD-ROM, or optical wirteable memories - and any other computer readable memory technologies.

Also, although this apparatus has been described primarily with reference to QEA™ analysis of human tissue
30 samples, the laboratory instruments and associated control, design, and analysis computer systems are not so limited. They are also adaptable to performing the CC embodiment of this invention and to the analysis of other samples, such as from animal models or in vitro cultures.
35   The invention is further described in the following exampl s which are in no way intended to limit the scop of the inventi n.

- 170 -

6.   EXAMPLES

6.1.   SUBSEQUENCE HIT AND LENGTH INFORMATION

This example illustrates QEA™ signals generated by
5   a PCR embodiment. From the October 1994 GenBank database,
12,000 human first continuous coding domain sequences ("CDS")
were selected. This selection resulted in a selected
database of sequences with a bias toward shorter genes, the
average length of the selected CDSs being 1000 bp instead of
10   the typical coding sequence length of 1800-2000 bp, and with
no guarantee that sequences were not be repeated in the
selected database. From this database, tables containing the
probability of occurrence of all 4 to 6-mer sequences were
constructed.

15   Then Eqns. 1 and 2 were solved for N = 12,000 and L
= 1,000 resulting in p = 0.17 and M = 108. Five 6-mer target
subsequences with this probability of occurrence were chosen
from the 6-mer tables and grouped into four pairs: CAGATA-
TCTCAC, CAGATA-GGTCTG, CAGATA-GCTCAA CAGATA-CACACC. Analyses
20   comprising mock digestions (see Sec. 5.4.1) of the selected
database of CDSs were then performed for these four pairs of
target subsequences.

The histogram of Fig. 1 presents the results of
these analyses. Along axis 102 is the length of fragments,
25   as would be observed in a gel separation of the amplified
fragments of a QEA™ reaction recognizing these target
subsequences. Along axis 101 is the number of fragments at a
given length. For example, spike 103 at a length of
approximately 800 base pairs represents three fragments of
30   the same length. Multiple fragments at one length may occur
either because several CDSs have one target subsequence pair
spaced this length, because one CDS has several target
subsequence pairs spaced this length, because of redundancy
in the selected CDSs, or because signals of this length were
35   generated by more than one pair of target subsequences.
Spike 104 at a slightly longer length r pres nts a singl
fragm nts. This fragment is generated from a unique sequenc

and provides a unique indication of its presence in a cDNA
mixture, that is, this is a good sequence.

6.2.   RESTRICTION ENDONUCLEASES

5   Tables 1-4 list all palindromic 4-mer and 6-mer
potential RE recognition sequences. RE enzymes recognizing
each site, where known, are also listed, along with an
exemplary commercial supplier. Over 85% of possible
sequences spanning a wide range of occurrence probabilities
10 have a known RE recognizing and cutting within the sequence.
        The frequency of these sequences was determined, as
in example 6.1, in 12,000 human first continuous coding
domain sequences selected from the October 1994 GenBank
database. The tables are sorted in order of increasing
15 recognition occurrence probability. The bar in the
recognition sequence indicates the site in the recognition
sequence where the RE cuts.
        The following vendor abbreviations are used: New England
Biolabs (Beverly, MA) ("NEB"), Stratagene (La Jolla, CA),
20 Boehringer Mannheim (Indianapolis, IN) ("BM"), and Gibco BRL
division of Life Technologies (Gaithersburg, MD) ("BRL").

25

30

35

TABLE 1:   THE 4-MER RESTRICTION SITES

Recognition Sequence	CDS Frequency	RE	Overhang	Vendor
C|GCG	0.36	SelI	2	
C|TAG	0.44	MaeI	2	NEB
T|TAA	0.45	MseI	2	NEB
TATA	0.45			
GCG|C	0.50	HhaI	2	NEB
ATAT	0.50			
A|CGT	0.52	MaeII	2	BM
T|CGA	0.53	TaqI	2	NEB
|AATT	0.53	Tsp509I	4	NEB
C|CGG	0.61	MspI	2	NEB
G|TAC	0.64	Csp6I	2	NEB
|GATC	0.67	Sau3AI	4	NEB
CATG|	0.68	NlaIII	4	NEB
TG|CA	0.78	CviRI	0	
AG|CT	0.78	AluI	0	NEB
GG|CC	0.79	HaeIII	0	NEB

TABLE 2: THE FIRST 20 6-MER RESTRICTION SITES

Sequence	CDS Frequency	RE	Overhang	Vendor
TCG|CGA	0.01	NruI	0	NEB
TAC|GTA	0.02	SnaBI	0	NEB
C|GTACG	0.02	BsiWI	4	NEB
CGAT|CG	0.02	PvuI	2	NEB
A|CGCGT	0.03	MluI	4	NEB
A|CTAGT	0.03	SpeI	4	NEB
G|TCGAC	0.04	SalI	4	NEB
AA|CGTT	0.04	Psp1406I	2	NEB
A|CCGGT	0.04	AgeI	4	NEB
G|CTAGC	0.04	NheI	4	NEB
TATATA	0.04			
GTT|AAC	0.05	HpaI	0	NEB
TAGCTA	0.05			
TAATTA	0.05			
GTA|TAC	0.05	Bst1107I	0	NEB
CTATAG	0.05			
CGCGCG	0.05			
C|CTAGG	0.06	AvrII	4	NEB
TT|CGAA	0.06	SfaI	2	BM
AT|CGAT	0.06	ClaI	2	NEB

TABLE 3: THE MIDDLE 20 6-MER RESTRICTION SITES

Sequence	CDS Frequency	RE	Overhang	Vendor
C|TTAAG	0.06	AflII	4	NEB
T|CTAGA	·0.06	XbaI	4	NEB
ATATAT	0.07			
AT|TAAT	0.07	VspI	2	BRL
G|CGCGC	0.08	BssHII	4	NEB
C|AATTG	0.08	MunI	4	NEB
GACGT|C	0.08	AatII	4	NEB
TTATAA	0.09			
TGC|GCA	0.10	FspI	0	NEB
C|TCGAG	0.01	XhoI	4	NEB
GAT|ATC	0.01	EcoRV	0	NEB
CA|TATG	0.10	NdeI	2	NEB
ATGCA|T	0.01	NsiI	4	NEB
AGC|GCT	0·11	Ecc47III	0	NEB
AAT|ATT	0.11	SspI	0	NEB
T|CCGGA	0.11	AccIII	4	Stratagene
TTT|AAA	0.12	DraI	0	NEB
A|CATGT	0.12	BspLVII	4	
CAC|GTG	0.12	Eco72I	0	Stratagene
CCGC|GG	0.12	SacII	2	NEB

5

10

15

20

25

30

35

TABLE 4: THE LAST 24 6-MER RESTRICTION SITES

Sequence	CDS Frequency	RE	Overhang	Vendor
GCATG|C	0.13	SphI	4	NEB
TTGCAA	0.13			
A|AGCTT	0.13	HindIII	4	NEB
G|TGCAC	0.13	ApaLI	4	NEB
AAATTT	0.14			
AGT|ACT	0.15	ScaI	0	NEB
G|AATTC	0.15	EcoRI	4	NEB
GGTAC|C	0.15	KpnI	4	NEB
T|GTACA	0.15	Bsp1407I	4	NEB
C|GGCCG	0.15	EagI	4	NEB
G|CCGGC	0.16	NgoMI	4	NEB
GGC|GCC	0.16	NarI	0	NEB
T|GATCA	0.16	BclI	4	NEB
T|CATGA	0.17	BspHI	4	NEB
C|CCGGG	0.19	SmaI	4	NEB
G|GATCC	0.19	BamHI	4	NEB
A|GATCT	0.20	BglII	4	NEB
AGG|CCT	0.22	StuI	0	NEB
GGGCC|C	0.24	ApaI	4	NEB
C|CATGG	0.24	NcoI	4	NEB
GAGCT|C	0.25	SacI	4	NEB
TGG|CCA	0.33	MscI	0	NEB
CAG|CTG	0.42	PvuII	0	NEB
CTGCA|G	0.43	PstI	4	NEB

6.3. RNA EXTRACTION AND cDNA SYNTHESIS

These protocols describe preferred methods for extraction of RNA from tissue samples and for synthesis of de-phosphorylated cDNA from the extracted RNA.

5

6.3.1. RNA EXTRACTION

In general, RNA extraction is done using Triazol reagent from Life Technologies (Gaithersburg, MD) following the protocol of Chomszynski et. al., 1987, *Annal. Biochem.*
10 162:156-59 and Chomszynski et. al., 1993, *Biotechniques* 15:532-34,536-37. Total RNA is first extracted from tissues, treated with Rnase-free Dnase I from Pharmacia Biotech (Uppsala, Sweden) to remove contaminating genomic DNA, followed by messenger RNA purification using oligo (dT)
15 magnetic beads from Dynal Corporation (Oslo, Norway), and then used for cDNA synthesis.

If desired, total cellular RNA can be separated into sub-pools prior to cDNA synthesis. For example, a sup-pool of endoplasmic reticulum associated RNA is enriched for
20 RNA producing proteins having an extra-cellular or receptor function.

In more detail, the following protocol is preferred for RNA extraction from tissue samples.

25 **Tissue Homogenization and Total RNA Extraction:**

A voxel is used to describe the specific piece of tissue to be analyzed. Most frequently it will refer to grid punches corresponding to pathologically characterized tissue sections.
30 1.   It is important that tissue voxels be quick frozen in liquid nitrogen immediately after dissection, and stored at -70°C until processed.
2.   The weight of the frozen tissue voxel is measured and recorded.
35 3.   Tissue voxels are pulverized and ground in liquid nitr gen, either with a porcelain mortar and pestle, or by stainless steel pulv rizers, or alternative means. This

tissue is ground to a fine powder and is kept on liquid
nitrogen.

4.   The tissue powder is transferred to a tube containing
Triazol reagent (Life Technologies, Gaithersburg, MD) with 1
5 ml of reagent per 100 mg of tissue and is dispersed in the
Triazol using a Polytron homogenizer from Brinkman
Instruments (Westbury, NY). For small tissue voxels less
than 100 mg, a minimum of 1 ml of Triazol reagent should be
used for efficient homogenization.

10 5.   Add 0.1 volumes BCP (1-bromo-3-chloropropane) (Molecular
Research, Cincinnati, OH) and mix by vortexing for 30
seconds. Let the mixture stand at room temperature for 15
minutes.

6.   Centrifuge for 15 minutes at 4°C at 12,000X G.

15 7.   Remove the aqueous phase to a fresh tube and add 0.5
volumes isopropanol per original amount of Triazol reagent
used and mix by vortexing for 30 seconds. Let the mixture
stand at room temperature for 10 minutes.

8.   Centrifuge at room temperature for 10 minutes at 12,000X
20 G.

9.   Wash with 70% ethanol and centrifuge at room temperature
for 5 minutes at 12,000X G.

10. Remove the supernatant and let the centrifuge tube stand
to dry in an inverted position.

25 11. Resuspend the RNA pellet in water (1 μl per mg of
original tissue weight) and heat to 55°C until completely
dissolved.

DNase treatment:

30 1.   Add 0.2 volume of 5X reverse transcriptase buffer (Life
Technologies, Gaithersburg, MD), 0.1 volumes of 0.1 M DTT,
and 5 units RNAguard per 100 mg starting tissue from
Pharmacia Biotech (Uppsala, Sweden).

2.   Add 1 unit RNase-free DNase I, Pharmacia Biotech, per
35 100 mg starting tissue. Incubate at 37°C for 20 minutes.

The following additional steps are optional,

- 178 -

Opt 1.   Repeat RNA extraction by adding 10 volumes f
           Triazol reagent.
Opt 2.   Repeat steps 5 through 11.

5 3.   Quantify the total RNA (from the RNA concentration
obtained by measuring OD_{260} of a 100 fold dilution). Store at
-20°C.

Isolation of Poly A⁺ Messenger RNA:

10   Poly-adenylated mRNA is isolated from total RNA
preparations using magnetic bead mediated oligo-dT detection.
Kits that can be used include Dynabeads mRNA Direct Kit from
Dynal (Oslo, Norway) or MPG Direct mRNA Purification Kit from
CPG (Lincoln Park, NJ). Protocols are used as directed by
15 the manufacturer.
          Less preferably, the following procedure can be
used. The Dynal oligo(dT) magnetic beads have a capacity of
1 ug poly(A⁺) per 100 ug of beads (1 mg/ml concentration),
assuming 2% of the total RNA has poly(A⁺) tails.
20

1.   Add 5 volumes of Lysis/Binding buffer (Dynal) and
sufficient beads to bind the estimated poly(A⁺) RNA.
2.   Incubate at 65°C for 2 minutes, then at room temperature
for 5 minutes.
25 3.   Wash beads with 1 ml Washing buffer/LiDS (Dynal)
4.   Wash beads with 1 ml Washing buffer (Dynal) 2 times.
5.   Elute poly(A⁺) RNA with 1 μl water/ug beads 2 times.

          For both methods, the poly-adenylated RNA is
30 harvested in a small volume of water, quantified as above,
and stored at -20°C. Typical yields of poly-adenylated RNA
range from 1% to 4% of the input total RNA.

6.3.2.   cDNA SYNTHESIS

35   This protocol for the synthesis of de-
phosphorylated cDNA from poly (A)+ RNA is pref rred when the

quantities of input RNA are approximately 1 μg, or at least 200 ng or gr ater.

Reagents Used:

- Random Hexamers (50 ng/μl)
- 5X First strand buffer (BRL)
- 10 mM dNTP mix
- 100 mM DTT
- SuperScript II reverse transcriptase (BRL) (200 U/μl)
- E. coli DNA ligase (BRL) 10 U/μl
- E. coli DNA polymerase (BRL) 10 U/μl
- T4 DNA polymerase 2.5 U/μl
- E. coli RNaseH (BRL) 3.5 U/μl
- Arctic Shrimp Alkaline Phosphatase, (SAP; USB), and 10X SAP buffer (USB)
- 5X Second strand buffer (BRL)
- 3 M Na-Acetate
- Phenol:Chloroform (phenol:chloroform:isoamyl alcohol 25:24:1)
- Chloroform isoamyl alcohol (24:1)
- Absolute and 75% ethanol
- 20 ug/μl glycogen (BM)

cDNA Synthesis Protocol:

1. Mix .25-1.0 ug of poly A+ RNA with 50 ng of random hexamers in 10 μl of water. Heat the mixture to 70°C for 10 min. and quick chill in ice-water slurry. Keep on ice for 1-2 min. Spin in microfuge for 10 secs. to collect condensate.

2. Prepare first stand reaction mix with 4 μl 5x First strand buffer, 2 μl 100 mM DTT, 1 μl 10 mM dNTP mix, and 2 μl water. Add this mix to the primer-annealed RNA from step 1. Place mixture at 37°C for 2 mins. Add 1 μl of Superscript II (BRL) (following manufacturer's recommendations). Incubate at 37°C for 1 hr.

3. Place tubes on ice, add 30 µl of 5x Second strand
    buffer, 90 µl of cold water, 3 µl of 10 mM dNTP, 1 µL
    (10 units) of *E. coli* DNA ligase, 4 µl (40 units) of *E.
    coli* DNA polymerase, and 1 µl (3.5 units) of *E. coli*

5   RnaseH. Incubate for 2h. at 16°C.

4. Add 2 µl of T4 DNA polymerase (5 units) and incubate at
    16°C for 5 min.

5. Add 20 µl 10x SAP buffer, 25 µl of water, and 5 µl (5
    units) of SAP. Incubate at 37°C for 30 min.

10 6. Extract cDNA with phenol-chloroform, chloroform-isoamyl
    alcohol. To the aqueous layer add Na-acetate to 0.3 M,
    20 ug glycogen, and 2 vol of ethanol. Incubate at -20°C
    for 10 min., spin at 14,000 g for 10 min. Wash pellet
    with 75% ethanol. Dissolve pellet in 50 µl TE.

15 7. Estimate the yield of cDNA using fluorometer.

8. For subsequent QEA™ processing, transfer 75 ng cDNA to a
    separate tube, add TE to make the concentration 600
    ng/ml and put that tube in the specified box at -20°C.
    For storage, add Na-acetate to 0.3M and 2 vol of ethanol

20 to the rest of cDNA and store at -80 °C.

    Alternative primers for first strand synthesis
known in the art can also be used for first strand synthesis.
Such primers include oligo(dT) primers, phasing primers, etc.
25

    ### 6.3.3. cDNA SYNTHESIS FOR SMALL QUANTITIES OF RNA
    The cDNA synthesis protocol previously described is
based primarily on the method of Gubler and Hoffman (Gubler
et al., 1983, "A simple and very efficient method for
30 generating cDNA libraries," *Gene* 25:263-9) and is robust and
well-proven for quantities of RNA in the 1 µg range (200 ng
and up). A more preferred protocol for RNA quantities below
200 ng takes advantage of the 5' CAP structure of RNAs (Edery
et al., 1995, "An efficient strategy to isolate full-length
35 cDNAs based on an mRNA cap retention procedure (CAPture),"
Mol. Cell Biol. 15:3363-71; Kato et al., 1994, "Construction

of a human full-length cDNA bank," *Gene* <u>150</u>:243-50). This
protocol has a number of advantages including:

   • broad scalability of RNA input quantities, making
them ideal for biopsies and for other small and variable
5 sized samples;

   • capability of doing a pre-QEA™ amplification of the
cDNA when very small amounts of cDNA are available;

   • cDNA synthesis biased toward full-length RNAs.

   • capability of introducing specific primer sites at
10 both ends of the full-length cDNAs;

   • option to eliminate the poly (A)+ RNA purification
step and use total RNA.

cDNA Synthesis Protocol

15 1.   The poly (A)$^+$, or total, RNA (10 μg) is dephosphorylated
        with bacterial alkaline phosphatase (20 μl rxn; 100 mM
        Tris-HCl pH 7.5, 2 mM DTT; 0.2 U bacterial alkaline
        phosphatase, 20 U Rnase inhibitor; 37°C for 30 minutes).

   2.   After phenol extraction and ethanol precipitation, the
20   RNA is treated with tobacco acid pyrophosphatase. (20
        μl rxn; 50 mM Na-OAc pH 6.0, 1 mM EDTA, 2 mM DTT; 0.1 U
        tobacco acid pyrophosphatase, 20 U Rnase inhibitor; 37°C
        for 30 minutes).

   3.   Phenol extract and ethanol precipitate the decapped RNA.
25   The following DNA-RNA primer named MA24R (3 nm) is
        ligated to the 5-prime end using T4 RNA Ligase (20 μl
        rxn; Tris-HCl pH 7.5, 5 mM MgCl$_2$, 0.5 mM ATP, 2 mM DTT,
        25% ethylene glycol; 100 U T4 RNA Ligase, 20 U Rnase
        inhibitor; 20°C for 12 hours):

30   MA24R: dCdAdGdTdAdGdCdGdAdTdTdGdCdCdGdCdCdGdTdCdAdGdGdTGGA
        (SEQ ID NO:??)

   4.   First strand synthesis is performed identically to steps
        1 and 2 of the protocol previously described in Sec.
        6.3.2 except that the following biotinylated primers are
35   used to prime the cDNA:

            MBTA: CGGTGGGTTGCCGTAGTAGCGGAT(T)$_{25}$A
            (SEQ ID NO:??)

MBTC: CGGTGGGTTGCCGTAGTAGCGGAT(T)$_{25}$C

(SEQ ID NO:??)

MBTG: CGGTGGGTTGCCGTAGTAGCGGAT(T)$_{25}$G

(SEQ ID NO:??)

5   These reactions can occur in separate tubes or in one
tube. The phasing effect of doing the reaction in
separate tubes has the advantage of dividing the cDNA
into three separate pools. 0.2 μg of each primer is
used in the reaction.

10 5.   Second strand synthesis is performed identically to
steps 3 and 4 of the protocol previously described in
Sec. 6.3.2 using a DNA-only version of the DNA-RNA
chimera is used to prime synthesis:

MA24:   CAGTAGCGATTGCCGCCGTCAGGT

15   (SEQ ID NO:??)

Because the primers at both 5' ends lack phosphate
groups, dephosphorylation of the resulting cDNA, e.g.,
by shrimp alkaline phosphatase, is no longer necessary.

6.   In cases where exceedingly small amounts of cDNA are
20   synthesized (1-10 ng yields), the sample can be
amplified using the following primer pair:

MA24:   CAGTAGCGATTGCCGCCGTCAGGT

(SEQ ID NO:??)

MB24:   CGGTGGGTTGCCGTAGTAGCGGAT

25   (SEQ ID NO:??)

For 1 ng quantities, 500-fold amplification by 8 to 10
PCR cycles (96°C 30 seconds, 57°C 1 minute, 72°C 3
minutes) provides adequate cDNA for comprehensive
analysis.

30

6.3.4.   ALTERNATIVE cDNA SYNTHESIS

cDNA synthesis

Alternately, cDNA can be synthesized using the
Superscript™ Choice system from Life Technologies, Inc.
35 (Gaithersburg, MD). If tissue voxels are the source for the
RNA, th polyadenylat d RNA is not quantified, and the entir
yield f polyad nylat d RNA is conc ntrated by precipitation

- 183 -

with ethanol. The polyadenylated RNA is resusp nded in 10 μl
of water, and 5 to 10 μl are us d for cDNA synthesis. The
manufacturer's protocols are followed for RNA amounts of less
than 1 μg, using 100 ng of random hexamers are used as
5 primers. If· greater than 1 μg of polyadenylated RNA is used,
the manufacturer's protocols are followed, using 50 ng of
random hexamer primers per microgram of polyadenylated RNA.
The resulting volume of the cDNA solution is 150 μl. If the
amount is not quantified, QEA™ test reactions can be run
10 using 1 μl or 0.1 μl of cDNA solution in order to determine
the appropriate amount of cDNA to use for subsequence QEA™
reactions.   '

cDNA De-phosphorylation

15   Where cDNA is synthesized with terminal phosphates,
they are preferably removed before the RE/Ligase reactions.
Terminal phosphate removal from cDNA is illustrated with the
use of Barents sea shrimp alkaline phosphatase ("SAP") (U.S.
Biochemical Corp.) and 2.5 μg of cDNA. Substantially less
20 (<10 ng) or more (>20 μg) of cDNA can be prepared at a time
with proportionally adjusted amounts of enzymes. Volumes are
maintained to preserve ease of handling. The quantities
necessary are consistent with using the method to analyze
small tissue samples from normal or diseased specimens.
25
    1.   Mix the following reagents
         2.5 μl   200 mM Tris-HCL
         23 μl   cDNA
         2 μl   2 units/μl Shrimp alkaline phosphatase
30   The final resulting cDNA concentration is 100 ng/μl.
    2.   Incubate at 37°C for 1 hour
    3.   Incubate at 80°C 15 minutes to inactivate the SAP.

6.4.   QEA™ PREFERRED RE METHOD

35   Protocols for the RE embodiment are designed to
minimize the number of individual manipulations down, and
thereby to maximize the reproducibility of QEA™ procedur s.

In preferred protocols, no buffer changes, precipitations, or organic (phenol/chloroform) extractions are used, all of which lower the overall efficiency of the process and reduce its utility for general use and more specifically for its use
5 in automated or robotic procedures.

Once the cDNA has been prepared, including terminal phosphate removal, it is separated into batches of from 1 ng to at least 50 ng each and of a number equal to the desired number of individual samples that need to be analyzed. For
10 example, if six RE/ligase reactions and six analyses are needed to generate all necessary signals, six batches are made. Advantageously, QEA™ reactions can be duplicated or triplicated in order to increase precision of and confidence in the results.

15   RE/ligase reactions are performed as digestions by, preferably, a pair of REs; alternatively, one or three or more REs can be used provided that the four base pair overhangs generated by each RE differ and can each be ligated to a uniquely adapter and that a sufficiently resolved length
20 distribution results. The preferably amount of RE enzyme specified in the protocols is sufficient for complete digestion while minimizing any other exo- or endo-nuclease activity that may be present in the enzyme. Preferred and alternate RE combinations can be found in Tables 11 to 14.

25   Adapters are chosen that are unique to each RE in a reaction. They are comprised of a linker complementary to each unique RE sticky overhang and a primer which uniquely hybridizes with that linker. The hybridized primer/linker combination is called an adapter.

30   The primer/linker combination for a giver RE are chosen according to the several embodiments of QEA™ reactions selected. Generally, sample primer/linker combinations are chosen according to the combinations illustrated in Table 10 for any particular RE. The primers can be labeled when the
35 detection means so require. Where one or more, or preferably all, primers have label m ieties, thes m ieties are pref rably distinguishable and can b advantage usly chos n

- 185 -

from the fluorescent labels described in Sec. 6.11. In a
QEA™ emb dim nt using post-PCR cleanup, one primer has a
capture moiety, e.g., biotin. The capture moiety is
preferably bound to one of the R-series primers, RA24 or
5 RC24, and the other primer is preferably labeled. Pairs of
labeled and biotinylated primers are preferably chosen
according to table 11 for the RE pairs therein listed.
Finally, in the case of an SEQ-QEA™ embodiment, primers and
linkers are preferably chosen according to Sec. 6.10.1.
10

6.4.1. PREFERRED RE/LIGASE & AMPLIFICATION REACTIONS .

This section describes the preferred protocol for
performing the RE/ligase and PCR amplification reactions with
a minimum of intervention.
15

Primer-excess Adapter Set Annealing

In the preferred protocol, a primer/linker
combination in the form of an adapter set specific for each
RE is chosen as above. The adapter set comprises sufficient
20 adaptors, hybridized primer/linker, for the RE/ligase
reaction and also sufficient excess primers for the
subsequent PCR amplification. Accordingly, primers do not
have to be separately added to the PCR reaction mix. Adapter
sets are constructed from linkers and primers according to
25 the following protocol:

1. Add to water linker and primer in a 1:20 concentration
ratio (12-mer : 24-mer) with the primer at a total
concentration of 50 pm per μl.
30 2. Incubate at 50°C for 10 minutes.
3. Cool slowly to room temperature and store at -20°C.

RE/ligase & Amplification Protocol

35 1. Combine the following components for the QPCR mix as
shown:

- 186 -

R ag nt	Conc ntration	1 rxn	96 rxns
10X TB 2.0	500 mM Tris pH 9.15, 160 mM $(NH_4)_2SO_4$, 20 mM $MgCl_2$	5 μl	525 μl
dNTP (equimolar mixture)	10 mM	2 μl	210 μl
Klentaq:PFU (16:1)	25 U/ml	0.25 μl	26.25 μl
water		32.75 μl	3438.75 μl
wax	90:10 Paraffin:Chillout™ 14.		

2.   Pre-wax PCR tubes by melting the 90:10
     Paraffin:Chillout™ 14 wax and adding the melted wax to
     the tubes in such a way that the wax solidifies on the
     sides of the upper half of the tubes.

3.   Mix solutions by tapping and/or inverting the tubes (do
     not vortex). Add 40μl QPCR mix to the pre-waxed PCR
     tubes. Add the solution one tube at a time carefully
     avoiding the sides and wax in the tubes. Note that it
     is important to keep the QPCR and the Qlig mixes
     separate as any QPCR mix in the ligation and the
     reaction will not work.

4.   The tubes are placed in a thermal cycler without lids
     and the wax is melted onto the liquid layer by
     incubating at 75°C for 2 min, followed by decreasing
     increments of 5°C for every 2 min until 25°C is reached.

5.   Combine the following components for the Qlig mix as
     shown:

R gent	C ncentrati n	1 rxn	24 rxn
RE 1	dep nds on RE	0.2µl	5.2µl
RE 2	depends on RE	0.2µl	5.2µl
Adapter set 1	20 pmole/ml for primer	1µl	26µl
Adapter set 2	20 pmole/ml for primer	1µl	26µl
ATP	10 mM	0.8µl	20.8µl
NEB 2	10X	1µl	26µl
Betaine	5 M	2µl	52µl..
Ligase	1 U/ml	0.2µl	5.2µl
H₂O		2.6µl	67.6µl

The amount for 24 rxns is advantageous for 8 cDNAs
reactions done in triplicate.

6.   After the Qlig mixes are complete for each set of
     enzymes the mix can be split up into tubes before adding
     the cDNAs. 24 reactions can be split up into 8 tubes
     each with 3 reaction volumes (approximately 27µl).
7.   Add the cDNA to the tubes and mix:

Reagent	Concentration	1 rxn	3 rxns
cDNA sample	1 ng/µl	1µl	3µl

The cDNA is prediluted to the appropriate concentration
of 1 ng/µl.

8.   Add 10µl of the Qlig mix to the top of the wax being
     careful not to disturb the wax. In the case where 24
     Qlig reactions are triplicated, the products can be
     split into 24 individual QPCR reactions.
9.   Gently add the caps to the tubes. Excess pressure can
     disturb the wax.
10.· Plac th tubes in a thermal cycler and perform the
     following thermal protocol.

Temp (in °C)	Time (in min.)	Reaction
37	30	Optimal RE digestion temperature
Ramp down to 37°C at -1°C/min.		
16	60	Optimal ligation temperature
37	15	Optimal RE digestion temperature
72	20	Melt wax; mix solutions in tube; blunt-end chains
Cycle the following steps for the number of PCR cycles, preferably 20		
96	30 sec.	Denaturing
57	1	Hybridizing
72	2	Chain elongation
End of the PCR cycles		
72	10	
4	hold	

11. After program is finished heat the tubes to 75°C for 5 minutes. Pull out the tubes and immediately turn them upside down till the wax hardens.

12. Place finished reactions in freezer or proceed directly to further processing.

The following are the preferred vendors for the various reagents used in this protocol.

R ag nts	V nd r	Catalog #
Enzymes	NEB (Beverly, MA)	
Adapters	Amitof/NBI (Allston, MA)	(see Table 10 for sequences)
Fluorescent Primers	Genosys (The Woodlands, TX)	(see Table 10 for sequences)
ATP	Pharmacia (Newark, NJ))	27-1006-02
dNTP	Pharmacia	27-2035-02
Klentaq	Ab peptides (St. Louis, MO)	1001
PFU	Stratagene (Los Angeles, CA)	600154
Betaine	Sigma (St. Louis, MO)	B-2754
Paraffin wax	Fluka Chemical, Inc. (Ronkonkoma, N.Y.)	76243
Chillout™ 14 liquid wax	MJ Research	
Ligase	BRL (Baltimore, MD)	15224-025

6.4.2. POST AMPLIFICATION CLEANUP PROTOCOL AND OTHER STEPS

Different post-amplification steps are appropriate for the various embodiments of QEA™/RE embodiment. In one case, QEA™ reaction are performed with labeled primers having no conjugated capture moieties. In this case, QEA™ reaction products are simply separated by length. When separation is by electrophoresis, the reaction products are suspended in a loading buffer and then loaded into an electrophoresis gel. A preferable electrophoresis apparatus is an ABI 377 (Applied Biosystems, Inc.) automated sequencer using the Gene Scan software (ABI) for analysis. The electrophoresis can be done under non-denaturing conditions, in which the dsDNA remains together and carries the labels (if any) of both primers. It can also be d n under denaturing conditi ns, in which each

- 190 -

ssDNA is separately lab led but typically are expected to migrate together.

In another case, one of the primers has a conjugated capture moiety, e.g., biotin, either for post-
5 amplification cleanup prior to separation or as part of the SEQ-QEA™ embodiment. In this case, QEA™ reaction products are first subject to a cleanup protocol for removing excess reagents and certain reaction products.

The following buffers are used in the post-PCR
10 cleanup protocol.

    Binding Buffer (H₂O solution)
        I.   5 M Nacl
        II. 10 mM Tris, pH 8.0
        III. 1 mM EDTA

    Wash Buffer (H₂O solution)
15   I.   10 mM Tris, pH 8.0
        II. 10 mM EDTA

    Loading Buffer (denaturing)
        I.   80% deionized formamide
        II. 20% 25 mM EDTA (pH 8.0), 50 mg/mL Blue
             dextran
20
    Ladder Loading Buffer
        I. 100 μL Gene Scan 500 ROX with 900 μL Loading
           Buffer

   Post-PCR Cleanup Protocol:
25 1.   Prepare enough streptavidin magnetic beads for purifying
        QEA™ products (Catalog No. MSTR0510 of CPG, Lincoln
        Park, N.J.). Use 3 μL of beads for every 5 μL of QEA™
        reaction product. Pre-wash beads in final suspension
        volume with binding buffer.

30

35

	1 Reaction		96 Reactions	
Sample Volume	Bead Volume	Suspension Volume	Bead Volume	Suspension Volume
5 μl	3 μl	10 μl	300 μl	1 ml
10 μl	6 μl	10 μl	600 μl	1 ml
15 μl	9 μl	10 μl	900 μl	1 ml
20 μl	12 μl	10 μl	1200 μl	1 ml

2. Dispense 10 μL of washed beads for every QEA™ sample to be processed. Purifications are done in a 96 well Falcon TC plate.

3. Add QEA™ product to beads. Mix well and incubate 30 minutes at 50°C.

4. Bring volume of sample up to 100 μl with binding buffer. Place plate on 96 well magnetic particle concentrator. Allow beads to migrate for 5 minutes.

5. Remove liquid, add 200 μL of washing buffer (TE pH 7.4).

6. Repeat the washing step 5.

7. In the case of a SEQ-QEA™ embodiment, the washed beads are now passed to the further steps of this embodiment as described in Sec. 6.5. In the other case of an embodiment using post-amplification cleanup alone, the washed beads are passed to the analysis step 9. Optionally, the beads may be stored by passing to step 10.

8. For analysis the beads are resuspended in loading buffer (5 μl for 5 μl of beads). Gene Scan 500 ROX ladder can be mixed in a one-tenth dilution. The supernatant is then analyzed by electrophoresis under denaturing conditions.

9. In case the beads are to be stored, remove liquid and air dry the beads.

10. Store plate dry in at -20°C.

In the case where on of primers has a conjugated
biotin moiety, QEA™ r action products fall into the following
three categories:

a)   A dsDNA molecule of which neither strand has a biotin
5   moiety;

b)   A dsDNA molecule having only one having with a
     conjugated biotin moiety;

c)   A dsDNA molecule having biotin moieties conjugated to
     both strands.

10 Category "a" products are not bound to the beads, and after
   the washing steps 5 and 6 of the previous protocol, they are
   washed from the beads, leaving only categories "b" and "c"
   attached to the beads. After step 9 in which the beads are
   resuspended in denaturing loading buffer, for category "b"
15 products, the strand not having the biotin moiety is released
   while the other strand with the biotin moiety is retained by
   the beads. For category "c" products, both strands are
   retained. Consequently, the electrophoresis of step 9
   separates single strands deriving from those reaction
20 products having only one conjugated biotin moiety.

6.4.3.   THE 5'-QEA™ EMBODIMENT

This subsection describes an exemplary protocol for
QEA™ embodiments which generates cDNA fragments which on the
25 5' end are fixed with respect to the 5' cap of the source
mRNA and which on the 3' end are singly cut by a chosen RE.
First, input cDNA is synthesized according to the protocol of
Sec. 6.3.3, or an equivalent protocol. Second, the protocols
in Sec. 6.4.1 and 6.4.2 previously described, except
30 differing only in the composition of the Qlig mix, are
performed.

1.   cDNA is synthesized according to the protocol of Sec.
     6.3.3.

35 2. The QPCR mix is prepared according to steps 1 through 4
     of th protocol of Sec. 6.4.1.

3.   Combine the following components for the Qlig mix as
     shown:

Regent	Concentration	1 rxn	24 rxn
RE 1	depends on RE	0.2µl	· 5.2µl
Adapter set 1	20 pmole/ml for primer	1µl	26µl
MA24 primer biotin-labeled	20 pmole/ml for primer	1µl	26µl
ATP	10 mM	0.8µl	20.8µl
NEB 2	10X	1µl	26µl
Betaine	5 M	2µl	52µl
Ligase	1 U/ml	0.2µl	5.2µl
H_2O		4.6µl	119.6µl

The amount for 24 reactions is advantageous for *
reactions performed in triplicate.

4.   The RE/ligase and PCR amplifications are processed
     according to steps 6 through 12 of the protocol of Sec.
     6.4.1.

5.   The reaction products are processed according to steps
     1-6 and 8-10 of the cleanup protocol of Sec. 6.4.2.

          After the washing step of the cleanup protocol,
step 6, attached to the streptavidin beads are only products
which are singly cut on the 3' end and are terminated at the
5' end by the biotin-labeled primer, which is ligated in a
fixed relation to the 5' cap of the source mRNA. Thus, upon
denaturing electrophoresis, step 9 of the cleanup protocol,
subsequent detection finds only signals from the desired
singly cut end fragments of definite length.

6.4.4.   FIRST ALTERNATIVE RE/LIGASE & AMPLIFICATION REACTIONS
          The section describes less preferred protocols
suitable for ith r manual or automated execution in two

tubes and suitable for labeled primers without a conjugated
capture moiety. Otherwise the REs and the primer/link r
components are chosen as previously described.

5 <u>Adapter Annealing</u>

    Adapters are former by annealing 12-mer linkers and
24-mer primers with some linker excess according to the
following protocol:

10 1.   Add to water linker and primer in a 2:1 concentration
    ratio (12-mer : 24-mer) with the primer at a total
    concentration of 5 pM per μl.
    2.   Incubate at 50°C for 10 minutes.
    3.   Cool slowly to room temperature and store at -20°C.
15

    Because there is no primer excess, primers must be
separately added to the PCR reaction mix.

<u>Manual RE/Ligase & Amplification Reactions</u>
20   This protocol is advantageously applied to separate
manually performed RE/Ligase and amplification reactions.
First, the RE/ligase reaction is prepared for use in a 96
well thermal cycler. Add per reaction:

25 1.   1 U of chosen REs (New England Biolabs, Beverly, MA)
        (preferred RE pair listing in Sec. 6.10)
    2.   1 μl of pre-annealed adapters appropriate for the chosen
        REs are prepared as above
    3.   1 μl of Ligase/ATP (0.2 μl T4 DNA ligase [1 U/μl]/0.8 μl
30   10 mM ATP from Life Technologies (Gaithersburg, MD))
    4.   0.5 μl 50 mM MgCl$_2$
    5.   10 ng of subject prepared cDNA
    6.   1 μl 10X NEB 2 buffer from New England Biolabs (Beverly,
        MA)
35 7.   Water to bring total volume to 10 μl

Th n perform the RE/ligation r action by following the th rmal profile in Fig. 16A using a PTC-100 Thermal Cycler from MJ Research (Watertown, MA).

Next for the PCR amplification reaction mix by
5 combining:

1.   10 μl 5X E-Mg (300 mM Tris-Hcl pH 9.0, 75 mM $(NH_4)_2SO_4$, no Mg ions))

2.   100 pm of appropriate fluorescently labeled 24-mer
10   primers

3.   1 μl 10 mM dNTP mix (Life Technologies, Gaithersburg, MD)

4.   2.5 U of 50:1 Taq polymerase (Life Technologies, Gaithersburg, MD) : Pfu polymerase (Stratagene, La
15   Jolla, CA)

5.   Water to bring volume to 40 μl per PCR reaction

Then perform the following steps:

20 1.   Add 40 μl of the PCR reaction mix to each RE/ligaticn reaction

2.   Perform the PCR temperature profile of Fig. 16B using a FTC-100 thermal cycler (MJ Research, Watertown, MA)

25

Automated RE/Ligase & Amplification Reactions

The preceding protocol can be advantageously automated according to the current protocol which requires intermediate reagent additions. Reactions are preformed in a
30 standard 96 well thermal cycler format using a Beckman Biomek 2000 robot (Beckman, Sunnyvale, CA). Typically 4 cDNA samples are analyzed in duplicate with 12 different RE pairs, for a total of 96 reactions. All steps are performed by the robot, including solution mixing, from user provided stock
35 reagents, and temperature profile control.

Pre-annealed adapters are pr par d as in the preceding section. Mix per RE/ligase reaction:

1.   1 U of appropriate RE (New England Biolabs, Beverly, MA)

2.   1 μl of appropriate annealed adapt r prepared as above
      (10 pm)

3.   0.1 μl T4 DNA ligase [1 U/μl] (Life Technologies
5   (Gaithersburg, MD)

4.   1 μl ATP (Life Technologies, Gaithersburg, MD)

5.   5 ng of subject prepared cDNA

6.   1.5 μl 10X NEB 2 buffer from New England Biolabs
      (Beverly, MA)

10 7. 0.5 μl of 50 mM $MgCl_2$

8.   Water to bring total volume to 10 μl and transfer to
      thermal cycler

        The robot requires 23 minutes total time to set up
15 the reactions. Then it performs the RE/ligation reaction by
following the temperature profile of Fig. 16C using a PTC-100
Thermal Cycler equipped with a mechanized lid from MJ
Research (Watertown, MA).

        Next, prepare the PCR reaction mix by combining:
20

1.   10 μl 5X E-Mg (300 mM Tris-HCl pH 9.0, 75 mM $(NH_4)_2SC_4$)

2.   100 pm of appropriate fluorescently labeled 24-mer
      primer

3.   1 μl 10 mM dNTP mix (Life Technologies, Gaithersburg,
25   MD)

4.   2.5 U of 50:1 Taq polymerase (Life Technologies,
      Gaithersburg, MD) : Pfu polymerase (Stratagene, La
      Jolla, CA)

5.   Water to being volume to 35 μl per PCR reaction
30

        Preheat the PCR mix to 72°C and transfer 35 μl of
the PCR mix to each digestion/ligation reaction and mix. The
robot requires 6 minutes for the transfer and mixing.
Perform the RE/ligase and PCR amplification reactions
35 according to the temperature profile of Fig. 16B using a PTC-
100 thermal cycl r equipped with a mechanized lid (MJ
Research, Wat rtown, MA).

The total elapsed time for the digestion/ligation
and PCR amplificati n reactions is 179 minutes. No user
intervention is required after initial experimental design
and reagent positioning.

5

   6.4.5.   SECOND ALTERNATIVE RE/LIG. & AMPLIFICATION REACTIONS
             The section describes a much less preferred fully
manual protocol in which the RE, the ligation, and the PCR
amplification reactions are all separately performed in three
10 tubes. It is suitable for labeled primers without a
conjugated capture moiety, with the REs and the primer/linker
components otherwise chosen as previously described. It is a
less preferred protocol.

15 **RE Digestion Reaction**
    1.   Mix the following reagents
          C.5 µl   prepared cDNA (100 ng/µl)) mixture (total 50 ng
                     of cDNA)
          10 µl   New England Biolabs Buffer No. 2
20   3 Units   RE enzyme
    2.   Incubate for 2 hours at 37°C.

          Larger size digests with higher concentrations of
cDNA can be used and fractions of the digest saved for
25 additional sets of experiments.

Adapter Ligation
          Since it is important to remove unwanted ligation
products, such as concatamers of fragments from different
30 cDNAs resulting from hybridization of RE sticky ends, the
restriction enzyme is left active during ligation. This
leads to a continuing cutting of unwanted concatamers and end
ligation of the desired end adapters. The majority of
restriction enzymes are active at the 16°C ligation
35 temperature. Ligation profiles consisting of optimum
ligation conditions int rsp rs d with optimum digestion
conditi ns can also be used to increase fficiency of this

process. An exemplary profile compris s p riodically cycling
between 37°C for 2 hours and 16°C for 2 hours at a ramp of
1°C/min.

One linker complementary to each 5 minutes overhang
5 generated by each RE is required. 100 picomoles ("pm") is a
sufficient molar excess for the protocol described. For each
linker a complementary uniquely labeled primer is added for
ligation to the cut ends of cDNAs. 100 pm is a sufficient
molar excess for the protocol described. If the amounts of
10 RE cDNA is changed the linker and primer amounts should be
proportionately changed.

Ligation Reaction (per 10 μl and 50 ng cDNA)
    1.   Mix the following reagents

15
    Component   Volume
    RE digested cDNA mixture   10 μl
    100 pM/μl each primer   1 μl
    100 pM/μl each linker   1 μl

20 2.   Thermally cycle from 50°C (a temperature at which the
          primers and linkers hybridize) to 10°C (-1°C/minute)
          then back to 16°C
    3.   Add 2 μl 10 mM ATP with 0.2 μl T4 DNA ligase (Premix
          0.1 μl ligase 1 U/μl per 1 μl ATP) (E. Coli ligase is a
25   less preferred alternative ligase.)
    4.   Incubate 12 hours at 16°C. This step can be shortened
          to less than 2 hours with proportionately higher ligase
          concentration. Alternately the thermal cycling protocol
          described can be used here.
30 5.   Incubate 2 hours 37°C
    6.   Incubate 20 minutes at 65°C to heat inactivate the
          ligase (last step should be RE cutting).
    7.   Hold at 4°C

35 Amplification Of Fragments With Ligated Adapters

This step amplifi s th fragm nts that have been
cut twice and ligated with adapt rs unique for ach RE cut
end. It is designed for a high amplification specificity.
Multiple amplifications are performed, with an increasing
5 number of amplification cycles. Use of the minimum number of
cycles to get the desired signal is preferred.
Amplifications above 20 cycles are not generally reliably
quantitative.

Mix the following to form the ligation mix:

10

Component	Volume
RE/Ligase cDNA mixture	5 μl
10X PCR Buffer	5 μl
25 mM MgCl$_2$	3 μl
10 mM dNTPs.	1 μl
100 pM/μl each primer	1 μl

15

Mix the following to form 150 μl PCR-Premix

20

Volume	Component
30 μl	Buffer E (ligation mix will contribute 0.3 mM MgCl)
1 μl	(300 pm/μl Rbuni24 Flour) 24 mer primer strand (50 pm/μl NBuni24 Tamra)
0.6 μl	Taq polymerase (per 150 μl)
3 μl	dNTP (10 mM)
106 μl	H$_2$O

25

Amplification of fragments is more specific if the
30 small linker dissociates from the ligated primer-cDNA complex
prior to amplification. The following is an exemplary method
for amplification of the results of six RE/ligase reactions.

1. Place three strips of six PCR tubes, marked 10, 15, and
35   20 cycles, into three rows on ice as shown.

20 cycles	1	2	3	4	5	6	-Add 140 μl PCR-
							premix
15 cycles	1	2	3	4	5	6	
5 10 cycles	1	2	3	4	5	6	-Add 10 μl
							ligation mix

2. Place 10 μl ligation mix in each tube in 10 cycle row

3. Place 140 μl PCR premix in each tube in 20 cycle row

10 4. Place into cycler and incubate for 5 minutes at 72°C.
   This melts linker which was not covalently ligated to the
   second strand of a cDNA fragment and allows the PCR
   premix to come to temperature.

5. Move the 140 μl PCR premix into the tubes in the 10 cycle
15   row containing the 10 μl ligation mix, then place 50 μl
   of result into corresponding tubes each in other rows.

6. Incubate for 5 minutes at 72°C. This finishes
   incompletely double stranded cDNA ends into complete
   dsDNA, the top primer being used as template for second
20   strand completion.

        The amplification cycle is designed to raise
specificity and reproducibility of the reaction. High
temperature and long melting times are used to reduce bias of
25 amplification due to high G+C content. Long extension times
are used to reduce bias in favor of smaller fragments. Long
denaturing times reduce PCR bias due to melting rates of
fragments, and long extension time reduces PCR bias on
fragment sizes.
30

1. Thermally cycle 95°C for 1 minute followed by 68°C for 3
   minutes.

2. Incubate at 72°C for 10 minutes at end of reaction.

35   6.4.6. OPTIONAL POST-AMPLIFICATION STEPS
        Sev ral optional steps can improve the signal from
the detected bands. First, single strands produced as a

- 201 -

result of lin ar amplification from singly cut fragments can
be remov d by the use of single strand specific exonuclease.
Exo I is the preferred single strand specific nuclease, and
is used by incubating 2 units of nuclease with the product of
5 each PCR reaction for 60 minutes at 37°C.

Second, the amplified products can be concentrated
prior to detection either by ethanol precipitation or column
separation with a hydroxyapatite column.

Several labeling methods are usable, including
10 fluorescent labeling as has been described, silver staining,
radiolabelled end primers, and intercalating dyes.
Fluorescent end labeling is preferred for high throughput
analysis with silver staining preferred if the individual
bands are to be removed from the gel for further processing,
15 such as sequencing.

Finally, fourth, use cf two primers allows direct
sequencing of separated strands by standard techniques. Also
separated strands can be directly cloned into vectors for use
in RNA assays such as in situ analysis. In that case, it is
20 more preferred to use primers containing T7 or other
polymerase signals.

25

30

35

6.5.   PREFERRED SEQ-QEA™ METHOD

The SEQ-QEA™ mbodim nt is practiced with the
special SEQ-QEA™ primers. One SEQ-QEA™ primer in a reaction
has a Type IIS restriction enzyme (e.g., FokI) recognition
5 site and a fluorescent tag, (e.g., FAM (carboxy-fluoroscein)
(ABI)) attached at the 5' end. The other primer used has a
biotin capture moiety ("Bio") and comprises either a uracil
residue or a site for a rare-cutting restriction enzyme like
AscI. Sec. 6.10.1 and Table 18 has a list of exemplary
10 primers and linkers for the SEQ-QEA™ methods.

Using these primers with corresponding linkers and
appropriate REs, the preferred QEA™ protocol of Sec. 6.4.1 is
performed and is followed by the post-PCR cleanup protocol of
Sec. 6.4.2 through the step 6 washing. As noted in step 7,
15 the products of step 6 are input to the further steps of the
SEQ-QEA™ embodiment.

The following are preferable primers and linkers to
be used together with the an RE1 of BglII and an RE2 of
EspHI.

20

SEQ-QEA™ method primer pairs	Type-IIS Enzyme	Method of Bead Release
1) KA5/KA24-FAM + RC9/UC24-Bio	FokI	UDG
2) BA5/BA24-FAM + RC9/UC24-Bio	BbvI	UDG
3) KA5/KA24-FAM + RC9/SC24-Bio	FokI	AscI
4) BA5/BA24-FAM + RC9/SC24-Bio	BbvI	AscI

25

Using the above REs and primer pairs, QEA™ method
reaction products obtained fall into the following three
categories:

a)   A double-stranded DNA with a 5' FAM label with nearby
30   sequence containing a recognition site for FokI or BbvI
on one strand, and a 3' biotin label with nearby sequence
containing a uracil residue or an AscI recognition site
on the other strand (in the case where different REs cut
at each end)

35 b)   A double-stranded DNA with a 5' biotin label with nearby
s quenc c ntaining a uracil r sidue or an AscI
recognition site on one strand, and a 3' biotin label

with nearby sequence containing a uracil residue or an
AscI recognition sit   n the other strand (in the case
where same RE cuts at both ends)

   c) A double-stranded DNA with a 5' FAM label with nearby

5   sequence containing a recognition site for FokI or BbvI
on one strand, and a 3' FAM label with nearby sequence
containing a recognition site for FokI or BbvI on the
other strand (in the case where same RE cuts at both
ends)

10   Typically, after QEA™ reactions according to the
protocol of Sec. 6.4.1 is completed, 45 μl out of 50 μl is
processed (the rest is saved). During the post-PCR cleanup
according to the protocol of Sec. 6.4.2, these 45 μl of the
reaction products are bound to the magnetic streptavidin

15 beads and washed at step 6 of this protocol. After this
step, only category "a" and "b" products are retained by th
magnetic streptavidin beads, the category "c" products having
no biotin moieties. Subsequently, the DNA bound to the beads
is digested with the Type IIS restriction enzyme in a volume

20 of 100 μl of a suitable 1X RE buffer, e.g. NEB 4 for FokI,
with about 10 units of the enzyme for 3 hours at 37°C. After
Type IIS RE digestion and washing only category "a" products
are retained by the beads, the category "b" products having
been cut at both ends and released from the beads. The

25 supernatant is then removed and the beads are washed three
times with the wash buffer. Type IIS restriction enzymes
cleave DNA at a location outside their recognition sites,
thus producing overhangs of unknown sequences (Szybalski et
al., 1991, Gene 100:13-26). The Type IIS digestion thus

30 releases the FAM label of the category "a" products and
creates a fragment-specific overhang that acts as a template
for sequencing. Complete Type IIS digestion can be checked
for by the absence of the FAM label.

        The end-sequencing reaction is essentially a chain

35 fill-in reaction using the overhang generated by the Type-IIS
restriction enzyme as a template. Dideoxy chain terminators
labeled with diff rent ABI fluorescent dyes are mixed at high

- 204 -

ratios with dNTPs to ensure high frequency of incorporation, and the DNA polym rase enzyme used (.g., Sequenas (T7 DNA polymerase), Taquenase (Taq polymerase)) has high affinity for the labeled dideoxynucleotides. A sequencing mix

5 totalling 20 μl containing the appropriate 1X buffer, 1 μl dNTPs diluted 1/200 from stock (3 mM dATP, 1.2 mM dCTP, 4.5 mM dGTP, 1.2 mM dTTP), 0.5 μl each ABI dye-labeled terminator solution (containing ddATP, ddCTP, ddGTP and ddTTP, respectively), (and 1 μl 0.1 M DTT for Sequenase) is

10 made. The beads are resuspended in the sequencing mix and 0.1 μl Taquenase is added and the reaction is incubated at 65°C for 15 minutes. If Sequenase is to be used, 0.1 μl Sequenase is added instead of Taquenase and the reaction is incubated at 37°C for 15 minutes. After this, the reaction

15 mix is transferred to a magnet and the supernatant is removed. The beads are washed twice with wash buffer f.

The above-described end-sequencing reaction incorporates dye labeled nucleotides into the strand that contains biotin. Since biotin-streptavidin binding is nearly

20 irreversible, the labeled strands must be cleaved for analysis by electrophoresis. This is achieved by treating UMP-containing fragments with Uracil DNA Glycosylase (UDG), or cleaving AscI-site-containing fragments with AscI. UDG removes the Uracil residue from dsDNA; the phosphate backbone

25 is subsequently hydrolyzed at temperatures above room temperature and at pH>8.3.

For UDG treatment, the beads are resuspended in 20 μl UDG buffer (30 mM Tris-HCl pH 7.5, 50 mM KCl, 5 mM MgCl$_2$), 0.2 units of UDG are added and the reaction is incubated at

30 room temperature for 30 minutes. The reaction is then transferred to a magnet and the supernatant removed. The biotinylated strand, which is the strand that is being filled in during end-sequencing, is still attached to the beads as UDG does not destroy the backbone, but makes it very

35 susceptible to hydrolysis.

Th beads ar resuspended in 5 μl f rmamide loading buffer. These are th n split into 2 tub s of 2.5 μl each.

- 205 -

Another 2.5 μl formamide loading buffer is added to one and
2.5 μl formamide loading buffer with 20% GS500 ROX ladder
(ABI) is added to the other. These are heated at 95°C for 5
minutes to effect hydrolysis and denaturation. Then they are
5   electrophoretically separated.

In case of the biotinylated primer having an AscI
site, the following is performed. The beads are resuspended
in 20 μl of AscI buffer (15 mM KOAc, 20 mM Tris, 10 mM MgOAc,
1 mM DTT, pH 7.9) and 5 units of AscI is added and incubated
10 at 37°C for 1 hour. The beads are separated on a magnet and
the supernatant that contains the digestion products is
precipitated with three volumes of ethanol after the addition
of 5 μg of glycogen. The pellet is resuspended in 5 μl
formamide loading buffer and split into 2 tubes of 2.5 μl
15 each. Another 2.5 μl formamide loading buffer is added to
one and 2.5 μl formamide loading buffer with 20% GS500 ROX
ladder is added to the other. These are heated at 95°C for 5
minutes and analyzed by electrophoretic separation.

Sequencing is completed by gel electrophoretic
20 separation of released and sequenced strands. The overhang
sequence is given by the order of partially filled in
fragments observed.

6.6. QEA™ BY THE PCR EMBODIMENT

25   This is an alternative QEA™ embodiment based on PCR
amplification of fragments between target subsequences
recognized by PCR primers or sets of PCR primers. It is
designed for the preferred primers described with reference
to Fig. 5. If other primers are used, such as simple sets of
30 degenerate oligonucleotides, step 5, the first low stringency
PCR cycle, is omitted.

First strand cDNA synthesis is carried out
according to Sec. 6.3. PCR amplification with defined sets
of primers is performed according to the following protocol.
35

1. Rnase treat the 1st strand mix with 1 μl of RNas
    Cocktail from Ambion, Inc. (Austin, TX) at 37°C for 30
    minutes.

2. Phenol/CHCl₃ extract the mixture 2 times, and purify it on

5   a Centricon 100, Milipore Corporation (Bedford, MA) using
    water as the filtrate.

3. Bring the end volume of the cDNA to 50 μl (starting with
    10 ng RNA/μl).

4. Set up the following PCR Reaction:

10

Component	Volume
cDNA (~10 ng/μl)	1 μl
10X PCR Buffer	2.5 μl
25 mM MgCl₂	1.5 μl
10 mM dNTPs	0.5 μl
20 pM/μl primer 1	2.5 μl
20 pM/μl primer 2	2.5 μl
Taq Poly. (5 U/μl)	0.2 μl
water	14.3 μl

15

20

5. One low stringency cycle with the profile:
            40°C for 3 minutes (annealing)
            72°C for 1 minute (extension)

6. Cycle using the following profile:

25   95°C for 1 minute
    15-30 times:
            95°C for 30 seconds
            50°C for 1 minute
            72°C for 1 minute

30   72°C for 5 minutes

7. 4°C hold.

8. Samples are precipitated, resuspended in denaturing
    loading buffer, and analyzed by electrophoresis (either
    under denaturing or non-denaturing conditions).

35

6.7.   EXAMPLE OF SIMULATED ANNEALING

From the October 1994 GenBank database containing human coding sequences, 12,000 of the first continuous coding domain sequences ("CDS") were selected as in Sec. 6.1. This

5 selection resulted in a selected database of sequences biased towards short sequences. Frequency tables were then created that listed the occurrence frequency of each nucleotide subsequence of lengths 4, 5, 6, 7, and 8 in this selected database. Test target subsequences were initially selected

10 whose probability of occurrence was near to 50%. This was feasible for the 4-mers, as they bind relatively frequently, but as the occurrence probability decreases with length, for longer sequences, the occurrence probability was often substantially less than 50%. These initially selected target

15 subsequences were then optimized, using the simulated annealing CC experimental design methods, to pick the best 16 subsequences.

Tables 5, 6 and 7 present the results for target subsequences of lengths 4, 5 and 6, respectively. Table 8

20 presents the results for optimizing target subsequences of length 4 through 6 together. Simulated annealing generally produced an approximately 20% improvement over a target subsequence selection guided only by the occurrence and independence probability criteria. This level of

25 optimization is likely to improve with larger and less redundant databases that represent longer genes. Longer sequences bind too infrequently in this database to make useful hash codes.

30   **TABLE 5:   AN OPTIMIZED SET OF 4-MER SUBSEQUENCES**

CGTC	GTTA	ACTA	CTAG
TTTT	TGTA	AATC	GTTG
TACC	TTGT	TTCG	GATA
CGGT	CTCG	AACG	GGTA

Th targ t subsequences in Table 5 were chosen from
all p ssible 256 4-mers. Ther ar 2.41 CDSs per hash code
on average. There was 692 CDSs (out of 12000) which are not
complementary to any of these subsequences.

5

TABLE 6:   AN OPTIMIZED SET OF 5-MER SUBSEQUENCES

AGGCA	ACTGT	GTCTC	TGTGC
CAACT	GCCCC	ACTAC	GTGAC
GCACC	GTCTG	GCCTC	CAGGT
AGGGG	GGAAC	GCTCC	GCTCT

The target subsequences in Table 6 were chosen from
the 300 most frequently occurring 5-mers. There are 2.33
15 CDSs per hash code on average.. There was 829 CDSs (out of
12000) which are not complementary to any of these
subsequences

TABLE 7:   AN OPTIMIZED SET OF 6-MER SUBSEQUENCES

20

TCCTCA	CCAGGC	AGCAGC	CTCCTG
AGCTGG	CTCTGG	CCAGGG	CAGAGA
GCCTGG	ACTGGA	CACCAT	GCTGTG
ACTGTG	TCTGTG	CCAAGG	CCTGGA

25

The target subsequences in Table 7 were chosen from
the 200 most frequently occurring 6-mers. There was 2.63
CDSs per hash code on average. There are 1530 CDSs (out of
12000) which are not complementary to any of these
30 subsequences.

35

TABLE 8:   AN PTIMIZED SET OF 4-, 5-, AND 6-MER SUBSEQUENCES

CTCG	TTCG	GATA	TTTT
CTAG	GGTA	ACTGT	ACTAC
CAACT	GTCTG	AGGCA	GCACC
TGTGC	GGAAC	AGGGG	CTCCTG

The target subsequences in Table 8 were chosen from sets in Tables 1-3. There was 2.22 CDSs per hash code on average. There are 715 CDSs (out of 12000) which are not complementary to any of these subsequences.

The bias of the selected CDSs toward short sequences, on the average less than the length of a typical gene, partially explains the 5-10% of CDSs that were not complementary to any selected target subsequence. Longer sequences would be expected to have more hits as they have more variability. Also more target subsequences can be chosen to improve coverage. The 2.2 to 2.6 CDSs per individual hash code is partially explained by replication in the selected database. No attempt was make to insure each CDS is unique in the selected database.

6.8.   QEA™ RESULTS

This subsection present results from QEA™ experiments directed primarily to the query and tissue modes.

6.8.1.   QUERY MODE QEA™ RESULTS

The pattern of gene expression differs from tissue to tissue, and is modulated both during normal development and during the progression of many diseases, including cancer. Query mode QEA™ experiments were used to investigate differences in gene expression between normal, hyperplastic, and adenocarcinomatous glandular tissues. We had at our disposal voxels containing all three types of tissue, preserved in such a way that the adjacent tissue sections were available for later *in situ* hybridization. The following xperiments were carried out with normal,

- 210 -

hyperplastic, and adenocarcinomatous tissue, respectively, of
a particular gland.

RNA Extraction and cDNA Synthesis

5   Isolation of total RNA and poly(A)$^+$ RNA from
homogenized glandular tissue voxels was performed
substantially as described in Sec. 6.3.1.   cDNA was prepared
substantially as described in Sec. 6.3.4.

10 Quantitative Expression Analysis

          QEA™ reactions were performed by the preferred RE
embodiment substantially as described in Sec. 6.4.4. This
included the following steps.

15 Adapter Annealing

          Pairs of 12-base and 24-base primers were pre-
annealed at a ratio of 2:1 (12 mer : 24 mer) at a
concentration of 5 picomoles 24 mer per microliter in 1X NEB
2 buffer. For linker/primer hybridization, the
20 oligonucleotide mixture was heated to 50°C for 10 minutes,
and allowed to cool slowly to room temperature. For this
experiment, 10 picomoles of JC3 and 5 picomoles of JC24, and
10 picomoles of RC6 and 5 picomoles of RC24 were separately
pre-annealed. The sequences of JC3, JC24, RC6, and RC24 are
25 listed in Table 10 of Sec. 6.10, infra.

Restriction-Digestion/Ligation Reaction

          Reactions were prepared in for use in a 8-well
thermal cycler format. Glandular cDNA isolated from 10
30 separate voxels of tissue was cut with HindIII and NgoMI, and
pre-annealed linkers were ligated onto the 4 base 5'
overhangs that these enzymes generated. Added per each QEA™
reaction were:

35   1 Unit of HindIII (New England Biolabs, Beverly MA)
       1 Unit of NgoMI (New England Biolabs, Bev rly MA)
       1 μl of pre-ann aled JC3/JC24

1 μl of pre-ann aled RC6/RC24

1 μl Ligase/ATP (0.2 μl T4 DNA Ligase (1 Unit/μl) and
    0.8μl 10 mM ATP - Life Technologies, Gaithersburg
    MD)

5   0.5μl 50 mM MgCL$_2$

10 ng of glandular cDNA

1μl 10X NEB 2 Buffer (New England Biolabs, Beverly MA)

Total volume of 10μl with H$_2$O

10   The temperature profile of Fig. 16A was performed
using a PTC-100 Thermal Cycler (MJ Research, Watertown MA).

Amplification Reaction

      The products of the RE/ligation reaction were then
15 amplified using RC24 and JC24 primers. The PCR reaction mix
included:

10μl 5X E-Mg (300 mM Tris-HCL pH 9.0, 75 mM (NH$_4$)$_2$SO$_4$)

100 pm RC24

20   100 pm JC24

1 μl 10 mM dNTP mix (Life Technologies, Gaithersburg MD)

2.5 Units 50:1 Taq polymerase (Life Technologies,
    Gaithersburg MD): Pfu polymerase (Stratagene, La
    Jolla CA) mix

25   Total volume of 40μl with H$_2$O.

      40μl preheated PCR reaction mix was added to each
restriction-digestion/ligation reaction. The temperature
profile of Fig. 16B was performed using a PTC-100 Thermal
30 Cycler (MJ Research, Watertown MA).

QEA™ Analysis

      The reaction products were separated on a 5%
acrylamide sequencing gel, and detected by silver staining.
35 Lane-to-lane comparisons were made both by visual inspection
of the gel, and by comparing computer enhanced images
obtained from scanning th gel using standard computer

scanner equipment. On particular band of length X bp was differentially xpressed, being prominent in some samples but absent in others. This band was picked from the gel, PCR re-amplified, and sequenced.

5   QEA™ analysis was performed substantially as described in Sec. 5.4.1 using the CDS database constructed as described in Sec. 6.1. Four possible sequences in that database were found to be possible contributors to a fragment of Y bp (note that Y bp = X - 46 bp, where PCR primers add 46

10 bp to the fragment length), sequences A, B, C, and D. Analysis of the sequencing of the picked band confirmed that this DNA fragment was produced by sequence C, which is presently entered in GenBank. This result confirms the correct functioning of the integrated experimental and

15 analysis methods.

          Further, analysis of sequence C predicted that a second double-digest, using REs BspHI and BstYI, would yield a second, non-overlapping restriction fragment at Z bp in length (plus the 46 bp of ligated primers). A second QEA™

20 reaction was performed using these glandular cDNAs. The previously described experimental conditions were used, with the exception of substituting BspHI, BstYI, RA5/RA24 and JC9/JC24 for HindIII, NgoMI, JC3/JC24 and RC6/RC24 during the RE/ligation reaction and of substituting RA24 and JC24 during

25 amplification reaction. Analysis of the results of this second QEA™ experiment on silver-stained acrylamide gels, as above, revealed the presence of a band of the predicted size, Z + 46 bp, that was also differentially expressed in the same tissue samples as the X bp fragment. This results confirms

30 the correct functioning of the mock digest prediction methods coupled with subsequence actual experimental digest.

          Additional hybrid primers were designed to facilitate direct sequencing of QEA™ products and the direct generation of RNA probes for the in situ hybridization to the

35 original tissue sample. The M13-21 primer or the M13 reverse primer (in italics) w re fused to the first 23 nucleotides of

JC24 and RC24 (in bold), respectively, to allow direct
s quencing of the double-digest d QEA™ products.

    M13-21J + JA24:   5' *GGC GCG CCT GTA AAA CGA CGG CCA GTA*
5 CCG ACG TCG·ACT ATC CAT GAA G 3'   (SEQ ID NO:56)
    M13revR + RA24:   5' *AAA ACT GCA GGA AAC AGC TAT GAC CAG*
CAC TCT CCA GCC TCT CAC CGA 3'   (SEQ ID NO:57)

In order to enable direct generation of anti-sense RNA probes
10 for *in situ* hybridization, the phage T7 promotor (in
   italics) was fused to the first 23 nucleotides of JA24/JC24
   and RA24/RC24 (in bold).

   T7 + JA24:   5' *ACT TCG AAA TTA ATA CGA CTC ACT ATA GGG* ACC
15 GAC GTC GAC TAT CCA TGA AG 3'   (SEQ ID NO:58)
   T7 + RA24:   5' *ACT TCG AAA TTA ATA CGA CTC ACT ATA GGG* AGC
   ACT CTC CAG CCT CTC ACC GA 3'   (SEQ ID NO:59)

6.8.2.   TISSUE MODE QEA™ RESULTS
20 Isolation of Human Placental Lactogen using QEA™
          Lactogen is one of the most highly expressed genes
   in the human placenta and has a known sequence. The sequence
   of lactogen was retrieved from GenBank and mock digestion
   reactions were performed, substantially as described in §
25 5.4.1, with a wide selection of possible RE pairs. These
   mock digestions showed that digesting placental cDNA with the
   restriction enzymes BssHIII and XbaI yields a lactogen
   fragment of 166 bp in length.

30 RNA Extraction and cDNA Synthesis
          Isolation of total RNA and poly(A)$^+$ RNA from
   homogenized human placenta tissue was performed substantially
   as described in Sec. 6.3.1. cDNA was prepared substantially
   as described in Sec. 6.3.4.
35

Quantitative Expression Analysis

QEA™ reactions were p rf rm d by the preferred RE embodiment substantially as described in Sec. 6.4.3. This included the following steps.

5

Adapter Annealing

Pairs of 12-base and 24-base primers were pre-annealed at a ratio of 2:1 (12 mer : 24 mer) at a concentration of 5 picomoles 24 mer per microliter in 1X NEB
10 2 buffer. The oligonucleotide mixture was heated to 50°C for 10 minutes, and allowed to cool slowly to room temperature. For this experiment, 10 picomoles of RC8 and 5 picomoles of RC24, and 10 picomoles of JC7 and 5 picomoles of JC24 were separately pre-annealed. The sequences of RC8, RC24, JC7,
15 and JC24 are set forth in Table 10 of Sec. 6.10, infra.

Restriction-Digestion/Ligation Reaction

Reactions were prepared for use in a 8-well thermal cycler format. Placental cDNA was cut with BssHII and XbaI,
20 and pre-annealed adapters ligated onto the 4 base 5' overhangs that these enzymes generated. Added per reaction were:

```
            1 Unit of BssHII (New England Biolabs, Beverly MA)
25          1 Unit of XbaI (New England Biolabs, Beverly MA)
            1 µl of pre-annealed RC8/RC24
            1 µl of pre-annealed JC7/JC24
            1 µl Ligase/ATP (0.2µl T4 DNA Ligase (1 Unit/µl) and
                    0.8µl 10 mM ATP - Life Technologies, Gaithersburg
30                  MD)
            0.5µl 50 mM MgCl₂
            10 ng of placental cDNA
            1µl 10X NEB 2 Buffer (New England Biolabs, Beverly MA)
            Total volume of 10µl with H₂O.
35
```

The temperature profile of Fig. 16A was performed using a PTC-100 Thermal Cycler (MJ Research, Watertown MA).

Amplification Reaction

The products of the RE/ligati n reaction w re then
amplified using RC24 and JC24 primers (see Table 10, infra).
The PCR reaction mix included:

5

       10μl 5X E-Mg (300 mM Tris-HCl pH 9.0, 75 mM (NH₄)₂SO₄)
       100 pm RC24
       100 pm JC24
       1μl 10 mM dNTP mix (Life Technologies, Gaithersburg MD)

10   2.5 Units 50:1 Taq polymerase (Life Technologies,
              Gaithersburg MD): Pfu polymerase (Stratagene, La
              Jolla CA) mix.
       Total volume of 40μl with H₂O.

15   40μl preheated PCR reaction mix was added to each
restriction-digestion/ligation reaction. The temperature
profile of Fig. 16B was performed using a PTC-100 Thermal
Cycler (MJ Research, Watertown MA).

20 QEA™ Analysis

          The reaction products were separated on a 5%
acrylamide sequencing gel and detected by silver staining. A
prominent band of size 212 bp was seen. This was predicted
to correspond to the 166 bp lactogen BssHII-XbaI fragment,
25 with JC24 ligated to the BssHII site, and RC24 ligated to the
XbaI site. To prove that this band did indeed correspond to
lactogen, the 212 bp band was excised from the gel, re-
amplified using JC24 and RC24, and the fragment was
sequenced. Analysis of these sequencing results proved that
30 the fragment was from lactogen. Moreover, the lactogen
sequence ended at the expected 4 base remnant of the
restriction site, immediately followed by either JC24 (at the
BssHII end) or RC24 (at the XbaI end).
          This result confirmed the experimental design
35 methods of Sec. 5.4.2 applied to selection of a QEA™
experiment to identify certain sequences of interest, in this
case the human plac ntal lactogen s qu nc , in a tissue cDNA

sampl . These design methods resulted in the selection of an
exp riment which successfully identified the g ne int nded.

Further QEA™ experiments were done according to the
protocols of this section on human placental derived cDNA

5 with differing enzyme combinations. One unit of each enzyme
of the enzyme combinations listed in the first column of
Table 9 were used in the restriction-digestion/ligation
reaction protocol. Primers and linkers for each RE were
chosen according to Table 10, with one appropriate "J" series

10 linker and primer and one appropriate "R" series linker and
primer used in each reaction. The reaction products were
separated by electrophoresis on a 5% acrylamide gel and the
bands detected by silver staining. Fragments from bands with
the lengths listed in the second column of Table 9 were

15 removed from the gel and sequenced. Sequencing identified
the subsequences on the ends of the fragments and the precise
lengths of each fragment. Each subsequence was
characteristic of one of the REs used, confirming correct
action of the ligation and amplification protocols. The

20 third column of Table 9 lists end subsequences, with a "1"
indicating the recognition subsequence of RE "Enz1" and a "2"
indicating the recognition subsequence of RE "Enz2".
Multiple fragments with the same length but differing
recognition subsequence are placed in separate sub-rows in

25 Table 9.

Mock digest reactions, as described in Sec. 5.4.1,
were performed using the CDS database selected according to §
6.1. These mock digestion reactions searched this CDS
database for sequences having recognition subsequences for

30 the REs and such that the recognition subsequences are spaced
apart in order to produce the fragments with the lengths
listed. This search identified the database accession
numbers listed in the fourth column of Table 9. The gene
responsible for each accession number was determined from a

35 GenBank lookup and is listed in the fifth column of Table 9.
Table 9. Each such g ne and its accompanying accession
numbers is list d in a further sub-row. Multiple accession

numbers associated with one gene reflect th redundancy
pr sent in curr nt CDS DNA sequence databases.

        For all but one of the fragments recovered from the
gel, the sequence for the fragment corresponded to one of the
5 genes identified by the mock digestion reaction as causing
that fragment. This particular gene is indicated by
displaying the gene name in underscore and bold in the fifth
column of Table 9. That the gene determined by sequencing
the separated fragment matched the prediction of the database
10 search confirms the efficacy of the experimental protocols
and the computer implemented experimental analysis and
ambiguity resolution methods of Sec. 5.4.1 and Sec. 5.4.3 for
tissue mode QEA™. In fact, the mock digestion reactions
provide a simple way of identifying possible ambiguities in
15 DNA sequence databases.

TABLE 9:   PLACENTA GENE CALLS

RE Combinations (Enz1 & Enz2)	Fragment Length	End Sub-seq.	Database Acc. Numbers	Gene Causing Fragment
BglII & BspE1	97	1,1	X07767	cAMP-Dependent Protein Kinase
		1,2	J03278, M21616	PDGF Receptor
			D23660, L20868, X73974	**Ribosomal Protein L4**
		2,2	M74096	Long Chain Acyl-CoA Dehydrogenase
BamH1 & BspE1	112	1,2	L26914, M93718, M95296	Nitric Oxide Synthase
			L22453, M90054, X73460	**Ribosomal Protein L3a**
BglII & BspE1	115	1,2	M20496, X05256	**Cathepsin L**

RE Combinations (Enz1 & Enz2)	Fragment Length	End Sub-seq.	Database Acc. Numbers	Gene Causing Fragment
BglII & NgoM1	137	2,2	X55740	5'-Nucleotidase
	137	1,2	L18967	TRP2 Dopachrome Tautomerase
			L10386	Tranglutaminase E3
			S69231	Tyrosinease-Related Protein 2
			X56998, X56999	__Ubiquitin__
EcoR1 & Bcl1	139	1,2	U14967	__Ribosomal Protein L21__
Bcl1 & NgoM1	144	1,2	J02984	__Ribosomal Protein S15__
			U04683, X80391	Olfactory Receptor OR17-40
	144	2,2	L12700	Engrailed-2
BamH1 & BspE1	144	1,2	X97234	Ribosomal Protein L11
			X14362	C3B/C4B Receptor
EcoR1 & HindIII	146	1,2	M13932	__Ribosomal Protein S17__
BssHII & Xba1	166	1,2	J00118, V00573	__Lactogen__
Bcl1 & NgoM1	168	1,2	S56985, X63527	__Ribosomal Protein L19__
BamH1 & BspE1	173	1,1	S59493, U10323	Nuclear Factor NF45
		1,2	M20882, M23575, M31125, M33666, M34420, M37399, M69245, M93061	__Pregnancy Sp. Glycoprotein beta 1__

RE Combinations (Enz1 & Enz2)	Fragment Length	End Sub-seq.	Databas Acc. Numbers	Gen Causing Fragment
BglII & NgoM1	192	1,1	D26350	Inositol Triphosphatase Receptor
			L27711, L25876	Protein Phosphatase CIP2/KAP1
		1,2	D29992, L27624	**Tissue Factor Pathway Inhibitor 2**
BglII & Age1	215	1,2	M11353, M11354	**Histone H3.3**

6.9. COLONY CALLING

The colony calling embodiment comprises the principal steps of cDNA library filter construction, PNA hybridization, and detection of hybridization. Determination of the sequence in a sample is done by the previously described computer implemented CC experimental analysis methods.

cDNA library filter construction

This protocol comprises three steps: first, robotic picking of colonies into microtiter plates, second, PCR amplification of inserts, and third, spotting of amplified cDNA inserts onto filters.

1.   Colony picking –
a)   Libraries are plated out at a density of 1,000–10,000 colonies per 100 mm Petri dish and are picked using a robot into 384 well microtiter plates containing 50 µl of TB medium with the appropriate antibiotic. There are several commercially available robots to do this task.   The preferable robot is from the Washington University Human Genome Sequencing Center (St. Louis, MO).

b)   The picked coloni s are grown for 8 hours at 37°C, and are frozen for archiving.

2. PCR amplification -
PCR primer pairs designed for insert amplification are
5   dispensed with a standard 25 μl PCR mix into 96 well microtiter plates. A 96 prong transfer tool picks and transfers samples to provide amplification templates from the 384 well colony into the 96 well PCR mixes. A standard 25 cycle amplification protocol generates 100-
10   500 ng of insert DNA.

3. Spotting on filers -
The PCR products are pooled back into a 384 well format microtiter plates identical to the colony plates above. Spotting onto filters is a service performed by Research
15   Genetics (Huntsville, AL).

Alternatively, cDNA library filters may be obtained from commercial sources in certain cases.

20 <u>PNA hybridization and detection</u>
PNAs are commercially available from Perseptive Biosystems (Bedford, MA). The protocol below uses 8 dyes on 16 different degenerate sets of PNA 8-mers containing as common subsequences the optimized 6-mer subsequences from
25 Table 7. Thereby, complete classification and determination of expressed genes in a human tissue can be done with only 4 hybridizations generating a code of length 32. Actual conditions for stringency may vary depending on the PNA set used.
30

1. Hybridization -
A pool of 8 PNAs are used, labeled with 8 different fluorochromes made up at a concentration of 0.1 μg/ml in 10 mM Phosphate buffer, pH 7.0, 1X Denhardt's solution
35   (20 mg/ml Ficoll 400, polyvinylpyrollidone, and BSA). The arrayed filters are hybridiz d for 16 hrs at 25° C, and

wash d 3 times in the above buffer without PNAs at a
temperature which maximizes signal/noise.

2. Visualization -

A fluorescent detection system, such as used for DNA
5   analysis, can be used to distinguish the dyes, and thus
the PNAs, present at each filter hybridization position.
PNA presence or absence defines a code for each
hybridization position on the filter.

10   **6.10.   PREFERRED QEA™ ADAPTERS AND REs PAIRS**
         Table 10 lists preferred primer-linker pairs that
may be used as adapters for the preferred RE embodiment of
QEA™.   The primers listed cover possible double-digest RE
combinations involving the approximately 56 available REs
15 generating a 5' 4 bp overhang.   There are 40 such REs
available from New England Biolabs.   For each QEA™ double
digest reaction, one primer and one linker from the "R"
series corresponding to one of the pair of REs and one primer
and one linker from the "J" series corresponding to the other
20 of the pair of REs are used together.   This choice satisfies
the adapter characteristics previously described.   Two pairs
from the same series are not compatible during amplification.

TABLE 10:   SAMPLE ADAPTERS

Series	Adapter: Primer (longer strand) Linker (shorter strand) Notes: 'm' signifies an optional label or capture moiety.	RE
RA24	5' m-AGC ACT CTC CAG CCT CTC ACC GAA 3' (SEQ ID NO:1)	
RA1	3' AG TGG CTT TTAA (SEQ ID NO:2)	Tsp509 I Mfel EcoRI
RA5	3' AG TGG CTT GTAC (SEQ ID NO:3)	NcoI BspHI
RA6	3' AG TGG CTT GGCC (SEQ ID NO:4)	XmaI NgoMI BspEI

RA7 (SEQ ID NO:5)	3' AG TGG CTT GCGC	BssHII AscI
RA8 (SEQ ID NO:6)	3' AG TGG CTT GATC	AvrII NheI XbaI
RA9 (SEQ ID NO:7)	3' AG TGG CTT CTAG	DpnII BamHI BclI
RA10 (SEQ ID NO:8)	3' AG TGG CTT CGCG	KasI
RA11 (SEQ ID NO:9)	3' AG TGG CTT CCGG	EagI Bsp120 I NotI EaeI
RA12 (SEQ ID NO:10)	3' AG TGG CTT CATG	BsiWI Acc65I BsrGI
RA14 (SEQ ID NO:11)	3' AG TGG CTT AGCT	XhoI SalI
RA15 (SEQ ID NO:12)	3' AG TGG CTT ACGT	ApaLI
RA16 (SEQ ID NO:13)	3' AG TGG CTT AATT	AflII
RA17 (SEQ ID NO:14)	3' AG TGG CTT AGCA	BssSI
RC24	5' m-AGC ACT CTC CAG CCT CTC ACC GAC 3' (SEQ ID NO:15)	
RC1 (SEQ ID NO:16)	3' AG TCG CTG TTAA	Tsp509 I EcoRI ApoI
RC3 (SEQ ID NO:17)	3' AG TCG CTG TCGA	HindII I
RC5 (SEQ ID NO:18)	3' AG TCG CTG GTAC	BspHI
RC6 (SEQ ID NO:19)	3' AG TCG CTG GGCC	AgeI NgoMI BspEI SgrAI BsrFI BsaWI

RC7	(SEQ ID NO:20)	3' AG TCG CTG GCGC	MluI BssHII AscI
RC8	(SEQ ID NO:21)	3' AG TCG CTG GATC	SpeI NheI XbaI
RC9	(SEQ ID NO:22)	3' AG TCG CTG CTAG	DpnII BglII BamHI BclI BstYI
RC10	(SEQ ID NO:23)	3' AG TCG CTG CGCG	KasI
RC11	(SEQ ID NO:24)	3' AG TCG CTG CCGG	Bsp120 I NotI
RC12	(SEQ ID NO:25)	3' AG TCG CTG CATG	Acc56I BsrGI
RC14	(SEQ ID NO:26)	3' AG TCG CTG AGCT	SalI
RCJ5	(SEQ ID NO:27)	3' AG TCG CTG ACGT	Ppu10I ApaLI
JA24	5' m-ACC GAC GTC GAC TAT CCA TGA AGA 3' (SEQ ID NO:28)		
JA1	(SEQ ID NO:29)	3' GT ACT TCT TTAA	Tsp509 I Mfe1 EcoRI
JA5	(SEQ ID NO:30)	3' GT ACT TCT GTAC	NcoI BspHI
JA6	(SEQ ID NO:31)	3' GT ACT TCT GGCC	XmaI NgoMI BspEI
JA7	(SEQ ID NO:32)	3' GT ACT TCT GCGC	BssHII AscI
JA8	(SEQ ID NO:33)	3' GT ACT TCT GATC	AvrII NheI XbaI
JA9	(SEQ ID NO:34)	3' GT ACT TCT CTAG	DpnII BamHI BclI
JA10	(SEQ ID NO:35)	3' GT ACT TCT CGCG	KasI

JA11	(SEQ ID NO:36)	3' GT ACT TCT CCGG	EagI Bsp120 I NotI EaeI
JA12	(SEQ ID NO:37)	3' GT ACT TCT CATG	BsiWI Acc65I BsrGI
JA14	(SEQ ID NO:38)	3' GT ACT TCT AGCT	XhoI SalI
JA15	(SEQ ID NO:39)	3' GT ACT TCT ACGT	ApaLI
JA16	(SEQ ID NO:40)	3' GT ACT TCT AATT	AflII
JA17	(SEQ ID NO:41)	3' GT ACT TCT AGCA	BssSI
JC24	5' m-ACC GAC GTC GAC TAT CCA TGA AGC 3' (SEQ ID NO:42)		
JC1	(SEQ ID NO:43)	3' GT ACT TCG TTAA	Tsp509 I EcoRI ApoI
JC3	(SEQ ID NO:44)	3' GT ACT TCG TCGA	HindII I
JC5	(SEQ ID NO:45)	3' GT ACT TCG GTAC	BspHI
JC6	(SEQ ID NO:46)	3' GT ACT TCG GGCC	AgeI NgoMI BspEI SgrAI BsrFI BsaWI
JC7	(SEQ ID NO:47)	3' GT ACT TCG GCGC	MluI BssHII AscI
JC8	(SEQ ID NO:48)	3' GT ACT TCG GTAC	SpeI NheI XbaI
JC9	(SEQ ID NO:49)	3' GT ACT TCG CTAG	DpnII BglII BamHI BclI BstYI

JC10	(SEQ ID NO:50)	3' GT ACT TCG CGCG	KasI
JC11	(SEQ ID NO:51)	3' GT ACT TCG CCGG	Bsp120 I NotI
JC12	(SEQ ID NO:52)	3' GT ACT TCG CATG	Acc56I BsrGI
JC14	(SEQ ID NO:53)	3' GT ACT TCG AGCT	SalI
JC15	(SEQ ID NO:54)	3' GT ACT TCG ACGT	Ppu10I ApaLI

In the case where one of the primers is conjugated to a capture moiety, Table 11 RE pairs and the corresponding primer/linker combinations that have been tested. This table supplements Table 10. Biotin can be conjugated to primers by using standard phosphoramidite chemistry.

TABLE 11:   TESTED RE PAIRS AND BIOTINYLATED ADAPTERS

RE 1	RE 2	Adapter 1	Adapter 2
		Chose labeled primer JA24 or JC24 to match the linker according to Table 10	Chose biotinylated primer RA24 or RC24 to match the linker according to Table 10
BamHI	BspHI	JC9	RA5
BglII	BspHI	JA5	RC9
BglII	EcoRI	JC1	RC9
BglII	HindIII	JC3	RC9
BglII	BspEI	JC6	RC9
BglII	Ncol	JC9	RA5
BspEI	BspYI	JC6	RC9
BspEI	HindIII	JC6	RC3
BspHI	EcoRI	JA5	RA1
BspHI	HindIII	JC3	RA5
BstYI	EcoRI	JC1	RC9
Ec RI	HindIII	JC3	RA1

BAMHI	HindIII	JC9	RC3
BspEI	BspHI	JC6	RA5
BspEI	EcoRI	JC6	RA1
BspHI	BstYI	JA5	RC9
BspHI	NgoMI	JA5	·RC6
BstYI	HindIII	JC3	RC9
HindIII	NcoI	JC3	RA5
HindIII	NgoMI	JC3	RC6

Tables 12 and 13 list the RE combinations that have
been tested in QEA™ experiments on human placental and
glandular cDNAs samples. The preferred double digests are
those that give more than approximately 50 bands in the range
of 100 to 700 bp. Table 12 lists the preferred RE
combinations for human cDNA analyses.

**TABLE 12: PREFERRED RE COMBINATIONS FOR
HUMAN cDNA ANALYSIS**

Acc56I & HindIII	Acc65I & NgoMI	BamHI & EcoRI
BglII & HindIII	BglII & NgoMI	BsiWI & BspHI
BspHI & BstYI	BspHI & NgoMI	BsrGI & EcoRI
EagI & EcoRI	EagI & HindIII	EagI & NcoI
HindIII & NgoMI	NgoMI & NheI	NgoMI & SpeI·
BglII & BspHI	Bsp120I & NcoI	BssHII & NgoMI
EcoRI & HindIII	NgoMI & XbaI	

Table 13 lists other RE combinations tested and
that can be used for human cDNA analyses.

TABLE 13:   THER RE C MBINATIONS F R HUMAN cDNA ANALYSIS

AvrII & NgoMI	BamHI & Bsp120I	BamHI & BspHI
BamHI & NcoI	BclI & BspHI	BclI & NcoI
BglII & BspEI	BglII & EcoRI	BglII & NcoI
BssHII & BsrGI	BstYI & NcoI	BamHI & HindIII
BglII & Bsp120I	BspHI & HindIII	

Tables 14 and 15 list the RE combinations that have been tested in QEA™ experiments on mouse cDNA samples. The preferred double digests are those that give more than approximately 50 bands in the range of 100 to 700 bp. Table 14 lists the preferred RE combinations for mouse cDNA analyses.

TABLE 14:   PREFERRED RE COMBINATIONS FOR MOUSE cDNA ANALYSIS

Acc56I & HindIII	Acc65I & NgoMI	AscI & HindIII
AvrII & NgoMI	BamHI & BspHI	BamHI & HindIII
BamHI & NcoI	BclI & NcoI	BglII & BspHI
BglII & HindIII	BglII & NcoI	BglII & NgoMI
Bsp120I & NcoI	Acc65I & BspHI	BspHI & Bsp120I
BspHI & BsrGI	BspHI & EagI	BspHI & NgoMI ·
BspHI & NotI	BssHII & HindIII	BstYI & HindIII
HindIII & NcoI	HindIII & NgoMI	NcoI & NotI
NgoMI & NheI	NgoMI & SpeI	NgoMI & XbaI
BclI & HindIII		

Table 15 lists other RE combinations tested and that can be used for mouse cDNA analyses.

TABLE 15:   OTHER RE C MBINATIONS FOR MOUSE cDNA ANALYSIS

Acc65I & Nc I	BclI & BspHI	BsiWI & BspHI
BsiWI & NcoI	BspHI & HindIII	BsrGI & NcoI
BssHII & NgoMI	BstYI & BspHII	EagI & NcoI
HindIII & MluI		

Table 16 lists the data obtained from various RE combinations using mouse cDNA samples. The number of bands was observed from silver stained acrylamide separation gels.

TABLE 16:   MOUSE cDNA RE DIGESTION RESULTS

RE Combination	Number of Bands
Acc65I & HindIII	200
Acc65I & NgoMI	150
AscI & HindIII	100
AvrII & NgoMI	50
BamHI & BspHI	200
BamHI & HindIII	150
BamHI & NcoI	150
BclI & BspHI	5
BclI & HindIII	150
BclI & NcoI	50
BglII & BspHI	50
BglII & HindIII	150
BglII & NcoI	50
BglII & NgoMI	50
Bsp120I & NcoI	50
BspHI & Acc65I	150
BspHI & Bsp120I	50
BspHI & BsrGI	200
BspHI & EagI	150
BspHI & HindIII	0

RE Combination	Number of Bands
BspHI & NgoMI	150
BspHI & NotI	150
BsrGI & NcoI	10
BssHII & HindIII	100
BssHII & NgoMI	20
BstY: & BspHI	20
BstYI & HindIII	200
EagI & NcoI	10
HindIII & MluI	25
HindIII & NcoI	50
HindIII & NgoMI	150
NcoI & NotI	200
NgoMI & NheI	50
NgoMI & SpeI	200
NgoMI & XbaI	50
TOTAL # BANDS	3490

31 available REs that recognize a 6 bp recognition sequence and generate a 4 bp 5' overhang are: Acc65I, AflII, AgeI, ApaLI, ApoI, AscI, AvrI, BamHI, BclI, BglII, BsiWI, Bsp120I, BspEI, BspHI, BsrGI, BssHII, BstYI, EagI, EcoRI, HindIII, MfeI, MluI, NcoI, NgoMI, NheI, NotI, Ppu10I, SalI, SpeI, XbaI, and XhoI.

All of these enzymes have been tested in QEA™ protocols according to Sec. 6.4.4 with the exception of AflII. All were useable except for MfeI, Ppu10I, SalI, and XhoI. All the other 26 enzymes have been tested and are usable in the RE implementation of QEA™.

However certain pairs of these enzymes are less informative due to the fact that they produce identical overhangs, and thus their recognition sequences cannot be distinguish d by QEA™ adapters. Thes pairs are Acc65I and

(BsiWI or BsrGI); AgeI and (BspEI or NcoMI); ApoI and EcoRI;
AscI and (BssHII or MluI); AvrI and (NheI, SpeI, or XbaI);
BamHI and (BclI, BglII, or BstYI); BclI and (BgLII or BstYI);
BglII and BstYI; BsiWI and BsrGI; Bsp120I and EagI; BspEI and

5 NcoMI; BspHI and NcoI; BssHII and MluI; NheI and (SpeI or
XbaI); and SpeI and XbaI.

     Thus 301 RE pairs have been tested and are useable
in the RE embodiments of QEA™.

10   **6.10.1. PREFERRED SEQ-QEA™ ENZYMES AND ADAPTERS**

     Table 17 lists exemplary Type IIS REs adaptable to
SEQ-QEA™ embodiment and their important characteristics. For
each RE, the table lists the recognition sequence on each
strand of a dsDNA molecule and the distance in bp from that

15 recognition sequence to the location of strand cutting. Also
listed is the net overhang generated.

TABLE 17: SAMPLE TYPE IIS REs

Type IIS RE	Recog. Seqs.	Dist. to cutting site (bp)	Over-hang (bp)	Comment
FokI	5'-GGATG CCTAC	9 13	4	
HgaI	5'-GACGC CTGCG	5 10	5	
BbvI	5'-GCAGC CGTCG	8 12	4	
BsmFI	5'-GGGAC CCCTG	10 14	4	Lower recognition site specificity
BspMI	5'-ACCTGC TGGACG	4 8	4	
SfaNI	5'-GCATC CGTAG	5 9	4	

35   Table 18 lists exemplary primer and linker
combinations adaptable to a SEQ-QEA™ method. They satisfy
the previously d scrib d requirem nts on prim rs and linkers.

- 231 -

Exc pt for the indicated differences, th y are the same as
the prim rs and link rs of similar names in Table 10. RA24-U
and RC24-U have a 5' biotin capture moiety and a uracil
release means as indicated, and are adaptable to the same
5 linkers and REs as are RA24 and RC24 of Table 10. RA24-S and
RC24-S also have a 5' biotin capture moiety with a AscI
recognition site release means as indicated in bold and
underlining, and are adaptable to the same linkers and REs as
are RA24 and RC24 of Table 10. JA24-K has an internal FokI
10 recognition site as indicated and a 5' FAM label moiety (see
Table 19). The FokI recognition site is optimally placed to
be used with a RE producing a 4 bp overhang. Linkers KA5,
KA6, and KA9 corresponding to the indicated REs function with
this primer. JC24-B has an internal BbvI recognition site, a
15 5' FAM label, and functions with linkers BA5 and BA9. The
BbvI recognition site is also optimally placed to be used
with a RE producing a 4 bp overhang.

TABLE 18: SAMPLE ADAPTERS

Series	Adapter: Primer (longer strand) Linker (shorter strand) Notes: 'b' signifies a biotin moiety 'f' signifies a FAM label moiety	RE
RA24-U	5' b-AGC ACT CTC CAG CC<u>U</u> CTC ACC GAA 3' (SEQ ID NO:??)	
RA24-S	5' b-AGC ACT CT<u>G GCG CGC C</u>TC ACC GAA 3' (SEQ ID NO:??)	
RC24-U	5' b-AGC ACT CTC CAG CC<u>U</u> CTC ACC GAC 3' (SEQ ID NO:??)	
RC24-S	5' b-AGC ACT CT<u>G GCG CGC C</u>TC ACC GAC 3' (SEQ ID NO:??)	
JA24-K	5' f-ACC GAC GTC GAC TAT <u>GGA TG</u>A AGA 3' (SEQ ID NO:??)	FokI (9)

35

KA9	(SEQ ID NO:??)	3' CT ACT TCT CTAG	DpnII BglII BamHI BclI BstYI
KA5	(SEQ ID NO:??)	3' CT ACT TCT GTAC	NcoI BspHI
KA6	(SEQ ID NO:??)	3' CT ACT TCT GGCC	AgeI NgoMI BspEI SgrAI BsrFI BsaWI
JC24-B	5' f-ACC GAC GTC GAC TAT CGC AGC AGA 3' (SEQ ID NO:??)		BbvI (8)
BA9	(SEQ ID NO:??)	3' CG TCG TCT CTAG	DpnII BglII BamHI BclI BstYI
BA5	(SEQ ID NO:??)	3' CG TCG TCT GTAC	NcoI BspHI

6.11. FLUORESCENT LABELS

Fluorochromes labels that can be used in the methods of the present invention include the classic fluorochromes as well as more specialized fluorochromes. The classic fluorochromes include bimane, ethidium, europium (III) citrate, fluorescein, La Jolla blue, methylcoumarin, nitrobenzofuran, pyrene butyrate, rhodamine, terbium chelate, and tetramethylrhodamine. More specialized fluorochromes are listed in Table 19 along with their suppliers.

TABLE 19: FLORESCENT LABELS

Fluorochrome	Vendor	Absorption Maximum	Emission Maximum
Bodipy 493/503	Molecular Probes	493	503
Cy2	BDS	489	505
Bodipy FL	Molecular Probes	508	516

Flu rochrome	Vend r	Absorption Maximum	Emission Maximum
FTC	Molecular Probes	494	518
FluorX	BDS	494	520
FAM	Perkin-Elmer	495	535
Carboxy-rhodamine	Molecular Probes	519	543
EITC	Molecular Probes	522	543
Bodipy 530/550	Molecular Probes	530	550
JOE	Perkin-Elmer	525	557
HEX	Perkin-Elmer	529	560
Bodipy 542/563	Molecular Probes	542	563
Cy3	BDS	552	565
TRITC	Molecular Probes	547	572
LRB	Molecular Probes	556	576
Bodipy LMR	Molecular Probes	545	577
Tamra	Perkin-Elmer	552	580
Bodipy 576/589	Molecular Probes	576	589
Bodipy 581/591	Molecular Probes	581	591
Cy3.5	BDS	581	596
XRITC	Molecular Probes	570	596
ROX	Perkin-Elmer	550	610
Texas Red	Molecular Probes	589	615
Bodipy TR (618?)	Molecular Probes	596	625
Cy5	BDS	650	667
Cy5.5	BDS	678	703
DdCy5	Beckman	680	710
Cy7	BDS	443	767
DbCy7	Beckman	790	820

The suppliers listed in Table 19 are Molecular Probes (Eugene, OR), Biological Detection Systems ("BDS") (Pittsburgh, PA) and Perkin-Elmer (Norwalk, CT).

Means of utilizing these fluorochromes by attaching
5   them to particular nucleotide groups are described in Kricka et al., 1995, Molecular Probing, Blotting, and Sequencing, chap. 1, Academic Press, New York. Preferred methods of attachment are by an amino linker or phosophoramidite chemistry.
10

7. SPECIFIC EMBODIMENTS, CITATION OF REFERENCES

The present invention is not to be limited in scope by the specific embodiments described herein. Indeed, various modifications of the invention in addition to those
15 described herein will become apparent to those skilled in the art from the foregoing description and accompanying figures. Such modifications are intended to fall within the scope of the appended claims.

Various publications are cited herein, the
20 disclosures of which are incorporated by reference in their entireties.

25

30

35

SEQUENCE LISTING

(1) GENERAL INFORMATION:

    (i) APPLICANT: Rothberg, Jonathan
              Deem, Michael
              Simpson, John

    (ii) TITLE OF INVENTION: Method and Apparatus for Identifying,
           Classifying, or Quantifying DNA Sequences in a Sample
           Without Sequencing

    (iii) NUMBER OF SEQUENCES: 70

    (iv) CORRESPONDENCE ADDRESS:
        (A) ADDRESSEE: Pennie and Edmonds
        (B) STREET: 1155 Avenue of the Americas
        (C) CITY: New York
        (D) STATE: New York
        (E) COUNTRY: USA
        (F) ZIP: 10036-2711

    (v) COMPUTER READABLE FORM:
        (A) MEDIUM TYPE: Floppy disk
        (B) COMPUTER: IBM PC compatible
        (C) OPERATING SYSTEM: PC-DOS/MS-DOS
        (D) SOFTWARE: PatentIn Release #1.0, Version #1.30

    (vi) CURRENT APPLICATION DATA:
        (A) APPLICATION NUMBER: To be assigned
        (B) FILING DATE: 14-June-1996
        (C) CLASSIFICATION:

    (viii) ATTORNEY/AGENT INFORMATION:
        (A) NAME: Misrock, S. Leslie
        (B) REGISTRATION NUMBER: 18,072
        (C) REFERENCE/DOCKET NUMBER: 7934-033-999

    (ix) TELECOMMUNICATION INFORMATION:
        (A) TELEPHONE: (212) 790-9090
        (B) TELEFAX: (212) 869-8864
        (C) TELEX: 66441 PENNIE

(2) INFORMATION FOR SEQ ID NO:1:

    (i) SEQUENCE CHARACTERISTICS:
        (A) LENGTH: 24 base pairs
        (B) TYPE: nucleic acid
        (C) STRANDEDNESS: single
        (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: DNA

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:1:

AGCACTCTCC AGCCTCTCAC CGAA   24

(2) INFORMATION FOR SEQ ID NO:2:

    (i) SEQUENCE CHARACTERISTICS:
        (A) LENGTH: 12 base pairs

```
                    (B) TYPE: nucl ic acid
                    (C) STRANDEDNESS: single
                    (D) TOPOLOGY: linear

              (ii) MOLECULE TYPE: DNA


   5

              (xi) SEQUENCE DESCRIPTION: SEQ ID NO:2:

        AGTGGCTTTT AA                                              12

        (2) INFORMATION FOR SEQ ID NO:3:

              (i) SEQUENCE CHARACTERISTICS:
  10              (A) LENGTH: 12 base pairs
                  (B) TYPE: nucleic acid
                  (C) STRANDEDNESS: single
                  (D) TOPOLOGY: linear

              (ii) MOLECULE TYPE: DNA



  15
              (xi) SEQUENCE DESCRIPTION: SEQ ID NO:3:

        AGTGGCTTGT AC                                              12

        (2) INFORMATION FOR SEQ ID NO:4:

              (i) SEQUENCE CHARACTERISTICS:
                  (A) LENGTH: 12 base pairs
  20              (B) TYPE: nucleic acid
                  (C) STRANDEDNESS: single
                  (D) TOPOLOGY: linear

              (ii) MOLECULE TYPE: DNA



  25          (xi) SEQUENCE DESCRIPTION: SEQ ID NO:4:

        AGTGGCTTGG CC                                              12

        (2) INFORMATION FOR SEQ ID NO:5:

              (i) SEQUENCE CHARACTERISTICS:
                  (A) LENGTH: 12 base pairs
                  (B) TYPE: nucleic acid
  30              (C) STRANDEDNESS: single
                  (D) TOPOLOGY: linear

              (ii) MOLECULE TYPE: DNA



              (xi) SEQUENCE DESCRIPTION: SEQ ID NO:5:
  35
        AGTGGCTTGC GC                                              12

        (2) INFORMATION FOR SEQ ID NO:6:
```

```
        (i) SEQUENCE CHARACTERISTICS:
            (A) LENGTH: 12 base pairs
            (B) TYPE: nucleic acid
            (C) STRANDEDNESS: single
            (D) TOPOLOGY: linear

       (ii) MOLECULE TYPE: DNA
   5



       (xi) SEQUENCE DESCRIPTION: SEQ ID NO:6:

    AGTGGCTTGA TC                                                12

    (2) INFORMATION FOR SEQ ID NO:7:
  10
        (i) SEQUENCE CHARACTERISTICS:
            (A) LENGTH: 12 base pairs
            (B) TYPE: nucleic acid
            (C) STRANDEDNESS: single
            (D) TOPOLOGY: linear

       (ii) MOLECULE TYPE: DNA

  15



       (xi) SEQUENCE DESCRIPTION: SEQ ID NO:7:

    AGTGCCTTCT AG                                                12

    (2) INFORMATION FOR SEQ ID NO:8:

  20    (i) SEQUENCE CHARACTERISTICS:
            (A) LENGTH: 12 base pairs
            (B) TYPE: nucleic acid
            (C) STRANDEDNESS: single
            (D) TOPOLOGY: linear

       (ii) MOLECULE TYPE: DNA



  25

       (xi) SEQUENCE DESCRIPTION: SEQ ID NO:8:

    AGTGGCTTCG CG                                                12

    (2) INFORMATION FOR SEQ ID NO:9:

        (i) SEQUENCE CHARACTERISTICS:
  30        (A) LENGTH: 12 base pairs
            (B) TYPE: nucleic acid
            (C) STRANDEDNESS: single
            (D) TOPOLOGY: linear

       (ii) MOLECULE TYPE: DNA



  35
       (xi) SEQUENCE DESCRIPTION: SEQ ID NO:9:

    AGTGGCTTCC GG                                                12
```

(2) INFORMATION FOR SEQ ID NO:10:

    (i) SEQUENCE CHARACTERISTICS:
        (A) LENGTH: 12 base pairs
        (B) TYPE: nucleic acid
        (C) STRANDEDNESS: single
        (D) TOPOLOGY: linear

5

    (ii) MOLECULE TYPE: DNA

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:10:

AGTGGCTTCA TG   12

10
(2) INFORMATION FOR SEQ ID NO:11:

    (i) SEQUENCE CHARACTERISTICS:
        (A) LENGTH: 12 base pairs
        (B) TYPE: nucleic acid
        (C) STRANDEDNESS: single
        (D) TOPOLOGY: linear

15   (ii) MOLECULE TYPE: DNA

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:11:

AGTGGCTTAG CT   12

20 (2) INFORMATION FOR SEQ ID NO:12:

    (i) SEQUENCE CHARACTERISTICS:
        (A) LENGTH: 12 base pairs
        (B) TYPE: nucleic acid
        (C) STRANDEDNESS: single
        (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: DNA
25

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:12:

AGTGGCTTAC GT   12

(2) INFORMATION FOR SEQ ID NO:13:
30
    (i) SEQUENCE CHARACTERISTICS:
        (A) LENGTH: 12 base pairs
        (B) TYPE: nucleic acid
        (C) STRANDEDNESS: single
        (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: DNA

35

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:13:

- 239 -

AGTGGCTTAA TT   12

(2) INFORMATION FOR SEQ ID NO:14:

    (i) SEQUENCE CHARACTERISTICS:
        (A) LENGTH: 12 base pairs
        (B) TYPE: nucleic acid
        (C) STRANDEDNESS: single
        (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: DNA

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:14:

AGTGGCTTAG CA   12

(2) INFORMATION FOR SEQ ID NO:15:

    (i) SEQUENCE CHARACTERISTICS:
        (A) LENGTH: 24 base pairs
        (B) TYPE: nucleic acid
        (C) STRANDEDNESS: single
        (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: DNA

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:15:

AGCACTCTCC AGCCTCTCAC CGAC   24

(2) INFORMATION FOR SEQ ID NO:16:

    (i) SEQUENCE CHARACTERISTICS:
        (A) LENGTH: 12 base pairs
        (B) TYPE: nucleic acid
        (C) STRANDEDNESS: single
        (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: DNA

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:16:

AGTCGCTGTT AA   12

(2) INFORMATION FOR SEQ ID NO:17:

    (i) SEQUENCE CHARACTERISTICS:
        (A) LENGTH: 12 base pairs
        (B) TYPE: nucleic acid
        (C) STRANDEDNESS: single
        (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: DNA

        (xi) SEQUENCE DESCRIPTION: SEQ ID NO:17:

AGTCGCTGTC GA   12

        (2) INFORMATION FOR SEQ ID NO:18:

            (i) SEQUENCE CHARACTERISTICS:
 5   (A) LENGTH: 12 base pairs
                (B) TYPE: nucleic acid
                (C) STRANDEDNESS: single
                (D) TOPOLOGY: linear

            (ii) MOLECULE TYPE: DNA

10   (xi) SEQUENCE DESCRIPTION: SEQ ID NO:18:

AGTCGCTGGT AC   12

        (2) INFORMATION FOR SEQ ID NO:19:

            (i) SEQUENCE CHARACTERISTICS:
                (A) LENGTH: 12 base pairs
15   (B) TYPE: nucleic acid
                (C) STRANDEDNESS: single
                (D) TOPOLOGY: linear

            (ii) MOLECULE TYPE: DNA

20   (xi) SEQUENCE DESCRIPTION: SEQ ID NO:19:

AGTCGCTGGG CC   12

        (2) INFORMATION FOR SEQ ID NO:20:

            (i) SEQUENCE CHARACTERISTICS:
                (A) LENGTH: 12 base pairs
                (B) TYPE: nucleic acid
25   (C) STRANDEDNESS: single
                (D) TOPOLOGY: linear

            (ii) MOLECULE TYPE: DNA

        (xi) SEQUENCE DESCRIPTION: SEQ ID NO:20:
30
AGTCGCTGGC GC   12

        (2) INFORMATION FOR SEQ ID NO:21:

            (i) SEQUENCE CHARACTERISTICS:
                (A) LENGTH: 12 base pairs
                (B) TYPE: nucleic acid
                (C) STRANDEDNESS: single
35   (D) TOPOLOGY: linear

            (ii) MOLECULE TYPE: DNA

      (xi) SEQUENCE DESCRIPTION: SEQ ID NO:21:

AGTCGCTGGA TC   12

   (2) INFORMATION FOR SEQ ID NO:22:

5   (i) SEQUENCE CHARACTERISTICS:
              (A) LENGTH: 12 base pairs
              (B) TYPE: nucleic acid
              (C) STRANDEDNESS: single
              (D) TOPOLOGY: linear

          (ii) MOLECULE TYPE: DNA

10

      (xi) SEQUENCE DESCRIPTION: SEQ ID NO:22:

AGTCGCTGCT AG   ‾12

   (2) INFORMATION FOR SEQ ID NO:23:

          (i) SEQUENCE CHARACTERISTICS:
15   (A) LENGTH: 12 base pairs
              (B) TYPE: nucleic acid
              (C) STRANDEDNESS: single
              (D) TOPOLOGY: linear

          (ii) MOLECULE TYPE: DNA

20
      (xi) SEQUENCE DESCRIPTION: SEQ ID NO:23:

AGTCGCTGCG CG   12

   (2) INFORMATION FOR SEQ ID NO:24:

          (i) SEQUENCE CHARACTERISTICS:
              (A) LENGTH: 12 base pairs
25   (B) TYPE: nucleic acid
              (C) STRANDEDNESS: single
              (D) TOPOLOGY: linear

          (ii) MOLECULE TYPE: DNA

30   (xi) SEQUENCE DESCRIPTION: SEQ ID NO:24:

AGTCGCTGCC GG   12

   (2) INFORMATION FOR SEQ ID NO:25:

          (i) SEQUENCE CHARACTERISTICS:
              (A) LENGTH: 12 base pairs
              (B) TYPE: nucleic acid
35   (C) STRANDEDNESS: single
              (D) TOPOLOGY: linear

          (ii) MOLECULE TYPE: DNA

(xi) SEQUENCE DESCRIPTI N: SEQ ID NO:25:

AGTCGCTGCA TG   12

(2) INFORMATION FOR SEQ ID NO:26:

5

     (i) SEQUENCE CHARACTERISTICS:
        (A) LENGTH: 12 base pairs
        (B) TYPE: nucleic acid
        (C) STRANDEDNESS: single
        (D) TOPOLOGY: linear

     (ii) MOLECULE TYPE: DNA

10

     (xi) SEQUENCE DESCRIPTION: SEQ ID NO:26:

AGTCGCTGAG CT   12

(2) INFORMATION FOR SEQ ID NO:27:

15   (i) SEQUENCE CHARACTERISTICS:
        (A) LENGTH: 12 base pairs
        (B) TYPE: nucleic acid
        (C) STRANDEDNESS: single
        (D) TOPOLOGY: linear

     (ii) MOLECULE TYPE: DNA

20

     (xi) SEQUENCE DESCRIPTION: SEQ ID NO:27:

AGTCGCTGAC GT   12

(2) INFORMATION FOR SEQ ID NO:28:

     (i) SEQUENCE CHARACTERISTICS:
25   (A) LENGTH: 24 base pairs
        (B) TYPE: nucleic acid
        (C) STRANDEDNESS: single
        (D) TOPOLOGY: linear

     (ii) MOLECULE TYPE: DNA

30
     (xi) SEQUENCE DESCRIPTION: SEQ ID NO:28:

ACCGACGTCG ACTATCCATG AAGA   24

(2) INFORMATION FOR SEQ ID NO:29:

     (i) SEQUENCE CHARACTERISTICS:
        (A) LENGTH: 12 base pairs
35   (B) TYPE: nucleic acid
        (C) STRANDEDNESS: single
        (D) TOPOLOGY: linear

     (ii) MOLECULE TYPE: DNA

   (xi) SEQUENCE DESCRIPTION: SEQ ID NO:29:

  GTACTTCTTT AA   12

5 (2) INFORMATION FOR SEQ ID NO:30:

     (i) SEQUENCE CHARACTERISTICS:
       (A) LENGTH: 12 base pairs
       (B) TYPE: nucleic acid
       (C) STRANDEDNESS: single
       (D) TOPOLOGY: linear

     (ii) MOLECULE TYPE: DNA
10

     (xi) SEQUENCE DESCRIPTION: SEQ ID NO:30:

  GTACTTCTGT AC   12

  (2) INFORMATION FOR SEQ ID NO:31:
15
     (i) SEQUENCE CHARACTERISTICS:
       (A) LENGTH: 12 base pairs
       (B) TYPE: nucleic acid
       (C) STRANDEDNESS: single
       (D) TOPOLOGY: linear

     (ii) MOLECULE TYPE: DNA

20

     (xi) SEQUENCE DESCRIPTION: SEQ ID NO:31:

  GTACTTCTGG CC   12

  (2) INFORMATION FOR SEQ ID NO:32:

25   (i) SEQUENCE CHARACTERISTICS:
       (A) LENGTH: 12 base pairs
       (B) TYPE: nucleic acid
       (C) STRANDEDNESS: single
       (D) TOPOLOGY: linear

     (ii) MOLECULE TYPE: DNA

30

     (xi) SEQUENCE DESCRIPTION: SEQ ID NO:32:

  GTACTTCTGC GC   12

  (2) INFORMATION FOR SEQ ID NO:33:

     (i) SEQUENCE CHARACTERISTICS:
35   (A) LENGTH: 12 base pairs
       (B) TYPE: nucleic acid
       (C) STRANDEDNESS: single
       (D) TOPOLOGY: linear

           (ii) MOLECULE TYPE: DNA

           (xi) SEQUENCE DESCRIPTION: SEQ ID NO:33:

5 GTACTTCTGA TC .   12

    (2) INFORMATION FOR SEQ ID NO:34:

         (i) SEQUENCE CHARACTERISTICS:
            (A) LENGTH: 12 base pairs
            (B) TYPE: nucleic acid
            (C) STRANDEDNESS: single
            (D) TOPOLOGY: linear
10
         (ii) MOLECULE TYPE: DNA

           (xi) SEQUENCE DESCRIPTION: SEQ ID NO:34:

GTACTTCTCT AG   12
15
    (2) INFORMATION FOR SEQ ID NO:35:

         (i) SEQUENCE CHARACTERISTICS:
            (A) LENGTH: 12 base pairs
            (B) TYPE: nucleic acid
            (C) STRANDEDNESS: single
            (D) TOPOLOGY: linear

20   (ii) MOLECULE TYPE: DNA

           (xi) SEQUENCE DESCRIPTION: SEQ ID NO:35:

GTACTTCTCG CG   12

25 (2) INFORMATION FOR SEQ ID NO:36:

         (i) SEQUENCE CHARACTERISTICS:
            (A) LENGTH: 12 base pairs
            (B) TYPE: nucleic acid
            (C) STRANDEDNESS: single
            (D) TOPOLOGY: linear

         (ii) MOLECULE TYPE: DNA
30

           (xi) SEQUENCE DESCRIPTION: SEQ ID NO:36:

GTACTTCTCC GG   12

    (2) INFORMATION FOR SEQ ID NO:37:
35
         (i) SEQUENCE CHARACTERISTICS:
            (A) LENGTH: 12 base pairs
            (B) TYPE: nucleic acid
            (C) STRANDEDNESS: singl

```
                (D) TOPOLOGY: linear

        (ii) MOLECULE TYPE: DNA




5       (xi) SEQUENCE DESCRIPTION: SEQ ID NO:37:

    GTACTTCTCA TG                                                          12

    (2) INFORMATION FOR SEQ ID NO:38:

            (i) SEQUENCE CHARACTERISTICS:
                (A) LENGTH: 12 base pairs
                (B) TYPE: nucleic acid
10              (C) STRANDEDNESS: single
                (D) TOPOLOGY: linear

        (ii) MOLECULE TYPE: DNA




        (xi) SEQUENCE DESCRIPTION: SEQ ID NO:38:
15
    GTACTTCTAG CT                                                          12

    (2) INFORMATION FOR SEQ ID NO:39:

            (i) SEQUENCE CHARACTERISTICS:
                (A) LENGTH: 12 base pairs
                (B) TYPE: nucleic acid
                (C) STRANDEDNESS: single
20              (D) TOPOLOGY: linear

        (ii) MOLECULE TYPE: DNA




        (xi) SEQUENCE DESCRIPTION: SEQ ID NO:39:

25  GTACTTCTAC GT                                                          12

    (2) INFORMATION FOR SEQ ID NO:40:

            (i) SEQUENCE CHARACTERISTICS:
                (A) LENGTH: 12 base pairs
                (B) TYPE: nucleic acid
                (C) STRANDEDNESS: single
                (D) TOPOLOGY: linear
30
        (ii) MOLECULE TYPE: DNA




        (xi) SEQUENCE DESCRIPTION: SEQ ID NO:40:

    GTACTTCTAA TT                                                          12
35
    (2) INFORMATION FOR SEQ ID NO:41:

            (i) SEQUENCE CHARACTERISTICS:
                (A) LENGTH: 12 base pairs
```

- 246 -

```
                    (B) TYPE: nucleic acid
                    (C) STRANDEDNESS: single
                    (D) TOPOLOGY: linear

              (ii) MOLECULE TYPE: DNA
```

5

```
              (xi) SEQUENCE DESCRIPTION: SEQ ID NO:41:

        GTACTTCTAG CA                                                    12

        (2) INFORMATION FOR SEQ ID NO:42:

              (i) SEQUENCE CHARACTERISTICS:
```
10
```
                    (A) LENGTH: 24 base pairs
                    (B) TYPE: nucleic acid
                    (C) STRANDEDNESS: single
                    (D) TOPOLOGY: linear

              (ii) MOLECULE TYPE: DNA
```

15
```
              (xi) SEQUENCE DESCRIPTION: SEQ ID NO:42:.

        ACCGACGTCG ACTATCCATG AAGC                                      24

        (2) INFORMATION FOR SEQ ID NO:43:

              (i) SEQUENCE CHARACTERISTICS:
                    (A) LENGTH: 12 base pairs
```
20
```
                    (B) TYPE: nucleic acid
                    (C) STRANDEDNESS: single
                    (D) TOPOLOGY: linear

              (ii) MOLECULE TYPE: DNA
```

25
```
              (xi) SEQUENCE DESCRIPTION: SEQ ID NO:43:

        GTACTTCGTT AA                                                   12

        (2) INFORMATION FOR SEQ ID NO:44:

              (i) SEQUENCE CHARACTERISTICS:
                    (A) LENGTH: 12 base pairs
                    (B) TYPE: nucleic acid
```
30
```
                    (C) STRANDEDNESS: single
                    (D) TOPOLOGY: linear

              (ii) MOLECULE TYPE: DNA
```

```
              (xi) SEQUENCE DESCRIPTION: SEQ ID NO:44:
```
35
```
        GTACTTCGTC GA                                                   12

        (2) INFORMATION FOR SEQ ID NO:45:
```

           (i) SEQUENCE CHARACTERISTICS:
                (A) LENGTH: 12 base pairs
                (B) TYPE: nucleic acid
                (C) STRANDEDNESS: single
                (D) TOPOLOGY: linear

        (ii) MOLECULE TYPE: DNA
5

        (xi) SEQUENCE DESCRIPTION: SEQ ID NO:45:

    GTACTTCGGT AC   12

    (2) INFORMATION FOR SEQ ID NO:46:
10
           (i) SEQUENCE CHARACTERISTICS:
                (A) LENGTH: 12 base pairs
                (B) TYPE: nucleic acid
                (C) STRANDEDNESS: single
                (D) TOPOLOGY: linear

        (ii) MOLECULE TYPE: DNA

15

        (xi) SEQUENCE DESCRIPTION: SEQ ID NO:46:

    GTACTTCGGG CC   12

    (2) INFORMATION FOR SEQ ID NO:47:

20   (i) SEQUENCE CHARACTERISTICS:
                (A) LENGTH: 12 base pairs
                (B) TYPE: nucleic acid
                (C) STRANDEDNESS: single
                (D) TOPOLOGY: linear

        (ii) MOLECULE TYPE: DNA

25

        (xi) SEQUENCE DESCRIPTION: SEQ ID NO:47:

    GTACTTCGGC GC   12

    (2) INFORMATION FOR SEQ ID NO:48:

           (i) SEQUENCE CHARACTERISTICS:
30   (A) LENGTH: 12 base pairs
                (B) TYPE: nucleic acid
                (C) STRANDEDNESS: single
                (D) TOPOLOGY: linear

        (ii) MOLECULE TYPE: DNA

35
        (xi) SEQUENCE DESCRIPTION: SEQ ID NO:48:

    GTACTTCGGT AC   12

                              - 248 -

(2) INFORMATION FOR SEQ ID NO:49:

    (i) SEQUENCE CHARACTERISTICS:
        (A) LENGTH: 12 base pairs
        (B) TYPE: nucleic acid
        (C) STRANDEDNESS: single
        (D) TOPOLOGY: linear

5

    (ii) MOLECULE TYPE: DNA

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:49:

GTACTTCGCT AG   12

10

(2) INFORMATION FOR SEQ ID NO:50:

    (i) SEQUENCE CHARACTERISTICS:
        (A) LENGTH: 12 base pairs
        (B) TYPE: nucleic acid
        (C) STRANDEDNESS: single
        (D) TOPOLOGY: linear

15   (ii) MOLECULE TYPE: DNA

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:50:

GTACTTCGCG CG   12

20 (2) INFORMATION FOR SEQ ID NO:51:

    (i) SEQUENCE CHARACTERISTICS:
        (A) LENGTH: 12 base pairs
        (B) TYPE: nucleic acid
        (C) STRANDEDNESS: single
        (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: DNA

25

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:51:

GTACTTCGCC GG   12

(2) INFORMATION FOR SEQ ID NO:52:

30
    (i) SEQUENCE CHARACTERISTICS:
        (A) LENGTH: 12 base pairs
        (B) TYPE: nucleic acid
        (C) STRANDEDNESS: single
        (D) TOPOLOGY: linear

    - (ii) MOLECULE TYPE: DNA

35

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:52:

GTACTTCGCA TG   12

    (2) INFORMATION FOR SEQ ID NO:53:

            (i) SEQUENCE CHARACTERISTICS:
                (A) LENGTH: 12 base pairs
                (B) TYPE: nucleic acid
5   (C) STRANDEDNESS: single
                (D) TOPOLOGY: linear

        (ii) MOLECULE TYPE: DNA

            (xi) SEQUENCE DESCRIPTION: SEQ ID NO:53:
10
GTACTTCGAG CT   12

    (2) INFORMATION FOR SEQ ID NO:54:

            (i) SEQUENCE CHARACTERISTICS:
                (A) LENGTH: 12 base pairs
                (B) TYPE: nucleic acid
                (C) STRANDEDNESS: single
15   (D) TOPOLOGY: linear

        (ii) MOLECULE TYPE: DNA

            (xi) SEQUENCE DESCRIPTION: SEQ ID NO:54:

20 GTACTTCGAC GT   12

    (2) INFORMATION FOR SEQ ID NO:55:

            (i) SEQUENCE CHARACTERISTICS:
                (A) LENGTH: 28 base pairs
                (B) TYPE: nucleic acid
                (C) STRANDEDNESS: single
                (D) TOPOLOGY: linear
25
        (ii) MOLECULE TYPE: DNA

            (xi) SEQUENCE DESCRIPTION: SEQ ID NO:55:

    AGCACTCTCC AGCCTCTCAC CGAGCATG   28
30
    (2) INFORMATION FOR SEQ ID NO:56:

            (i) SEQUENCE CHARACTERISTICS:
                (A) LENGTH: 49 base pairs
                (B) TYPE: nucleic acid
                (C) STRANDEDNESS: single
                (D) TOPOLOGY: linear

35   (ii) MOLECULE TYPE: DNA

        (xi) SEQUENCE DESCRIPTION: SEQ ID NO:56:

    GGCGCGCCTG TAAAACGACG GCCAGTACCG ACGTCGACTA TCCATGAAG   49

        (2) INFORMATION FOR SEQ ID NO:57:

            (i) SEQUENCE CHARACTERISTICS:
5   (A) LENGTH: 48 base pairs
                 (B) TYPE: nucleic acid
                 (C) STRANDEDNESS: single
                 (D) TOPOLOGY: linear

        (ii) MOLECULE TYPE: DNA

10   (xi) SEQUENCE DESCRIPTION: SEQ ID NO:57:

    AAAACTGCAG GAAACAGCTA TGACCAGCAC TCTCCAGCCT CTCACCGA   48

        (2) INFORMATION FOR SEQ ID NO:58:

            (i) SEQUENCE CHARACTERISTICS:
                 (A) LENGTH: 53 base pairs
15   (B) TYPE: nucleic acid
                 (C) STRANDEDNESS: single
                 (D) TOPOLOGY: linear

        (ii) MOLECULE TYPE: DNA

20   (xi) SEQUENCE DESCRIPTION: SEQ ID NO:58:

    ACTTCGAAAT TAATACGACT CACTATAGGG ACCGACGTCG ACTATCCATG AAG   53

        (2) INFORMATION FOR SEQ ID NO:59:

            (i) SEQUENCE CHARACTERISTICS:
                 (A) LENGTH: 53 base pairs
                 (B) TYPE: nucleic acid
25   (C) STRANDEDNESS: single
                 (D) TOPOLOGY: linear

        (ii) MOLECULE TYPE: DNA

        (xi) SEQUENCE DESCRIPTION: SEQ ID NO:59:
30
    ACTTCGAAAT TAATACGACT CACTATAGGG AGCACTCTCC AGCCTCTCAC CGA   53

        (2) INFORMATION FOR SEQ ID NO:60:

            (i) SEQUENCE CHARACTERISTICS:
                 (A) LENGTH: 24 base pairs
                 (B) TYPE: nucleic acid
                 (C) STRANDEDNESS: single
35   (D) TOPOLOGY: linear

        (ii) MOLECULE TYPE: DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:60:

AGCACTCTCC AGCCUCTCAC CGAA   24

(2) INFORMATION FOR SEQ ID NO:61:

5   (i) SEQUENCE CHARACTERISTICS:
               (A) LENGTH: 24 base pairs
               (B) TYPE: nucleic acid
               (C) STRANDEDNESS: single
               (D) TOPOLOGY: linear

          (ii) MOLECULE TYPE: DNA

10

          (xi) SEQUENCE DESCRIPTION: SEQ ID NO:61:

AGCACTCTGG CGCGCCTCAC CGAA   24

(2) INFORMATION FOR SEQ ID NO:62:

          (i) SEQUENCE CHARACTERISTICS:
15   (A) LENGTH: 24 base pairs
               (B) TYPE: nucleic acid
               (C) STRANDEDNESS: single
               (D) TOPOLOGY: linear

          (ii) MOLECULE TYPE: DNA

20
          (xi) SEQUENCE DESCRIPTION: SEQ ID NO:62:

AGCACTCTCC AGCCUCTCAC CGAC   24

(2) INFORMATION FOR SEQ ID NO:63:

          (i) SEQUENCE CHARACTERISTICS:
               (A) LENGTH: 24 base pairs
25   (B) TYPE: nucleic acid
               (C) STRANDEDNESS: single
               (D) TOPOLOGY: linear

          (ii) MOLECULE TYPE: DNA

30   (xi) SEQUENCE DESCRIPTION: SEQ ID NO:63:

AGCACTCTGG CGCGCCTCAC CGAC   24

(2) INFORMATION FOR SEQ ID NO:64:

          (i) SEQUENCE CHARACTERISTICS:
               (A) LENGTH: 24 base pairs
               (B) TYPE: nucleic acid
35   (C) STRANDEDNESS: single
               (D) TOPOLOGY: linear

          (ii) MOLECULE TYPE: DNA

      (xi) SEQUENCE DESCRIPTION: SEQ ID NO:64:

ACCGACGTCG ACTATGGATG AAGA   24

  (2) INFORMATION FOR SEQ ID NO:65:

5

      (i) SEQUENCE CHARACTERISTICS:
          (A) LENGTH: 12 base pairs
          (B) TYPE: nucleic acid
          (C) STRANDEDNESS: single
          (D) TOPOLOGY: linear

      (ii) MOLECULE TYPE: DNA

10

      (xi) SEQUENCE DESCRIPTION: SEQ ID NO:65:

GATCTCTTCA TC   12

  (2) INFORMATION FOR SEQ ID NO:66:

15   (i) SEQUENCE CHARACTERISTICS:
          (A) LENGTH: 12 base pairs
          (B) TYPE: nucleic acid
          (C) STRANDEDNESS: single
          (D) TOPOLOGY: linear

      (ii) MOLECULE TYPE: DNA

20

      (xi) SEQUENCE DESCRIPTION: SEQ ID NO:66:

CATGTCTTCA TC   12

  (2) INFORMATION FOR SEQ ID NO:67:

      (i) SEQUENCE CHARACTERISTICS:
25   (A) LENGTH: 12 base pairs
          (B) TYPE: nucleic acid
          (C) STRANDEDNESS: single
          (D) TOPOLOGY: linear

      (ii) MOLECULE TYPE: DNA

30
      (xi) SEQUENCE DESCRIPTION: SEQ ID NO:67:

CCGGTCTTCA TC   12

  (2) INFORMATION FOR SEQ ID NO:68:

      (i) SEQUENCE CHARACTERISTICS:
         (A) LENGTH: 24 base pairs
35   (B) TYPE: nucleic acid
         (C) STRANDEDNESS: single
         (D) TOPOLOGY: linear

      (ii) MOLECULE TYPE: DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:68:

ACCGACGTCG ACTATCGCAG CAGA   24

5 (2) INFORMATION FOR SEQ ID NO:69:

    (i) SEQUENCE CHARACTERISTICS:
       (A) LENGTH: 12 base pairs
       (B) TYPE: nucleic acid
       (C) STRANDEDNESS: single
       (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: DNA
10

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:69:

GATCTCTGCT GC   12

(2) INFORMATION FOR SEQ ID NO:70:
15
    (i) SEQUENCE CHARACTERISTICS:
       (A) LENGTH: 12 base pairs
       (B) TYPE: nucleic acid
       (C) STRANDEDNESS: single
       (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: DNA

20

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:70:

CATGTCTGCT GC   12

25

30

35

WHAT IS CLAIMED IS:

1.   A method for identifying, classifying, or
quantifying one or more nucleic acids in a sample comprising
5 a plurality of nucleic acids having different nucleotide
sequences, said method comprising:
(a) probing said sample with one or more recognition
means, each recognition means recognizing a different target
nucleotide subsequence or a different set of target
10 nucleotide subsequences;
(b) generating one or more signals from said sample
probed by said recognition means, each generated signal
arising from a nucleic acid in said sample and comprising a
representation of (i) the length between occurrences of
15 target subsequences in said nucleic acid, and (ii) the
identities of said target subsequences in said nucleic acid
or the identities of said sets of target subsequences among
which are included the target subsequences in said nucleic
acid; and
20   (c) searching a nucleotide sequence database to
determine sequences that match or the absence of any
sequences that match said one or more generated signals, said
database comprising a plurality of known nucleotide sequences
of nucleic acids that may be present in the sample, a
25 sequence from said database matching a generated signal when
the sequence from said database has both (i) the same length
between occurrences of target subsequences as is represented
by the generated signal, and (ii) the same target
subsequences as are represented by the generated signal, or
30 target subsequences that are members of the same sets of
target subsequences represented by the generated signal,
whereby said one or more nucleic acids in said sample
are identified, classified, or quantified.

35   2.   The method of claim 1 wherein each r cognition
means recognizes on target subsequence, and wherein a
sequence from said databas matches a g n rat d signal when

the sequence fr m said database has both the same length
between occurrenc s of target subsequences as is represented
by the generated signal and the same target subsequences as
represented by the generated signal.

5

        3.   The method of claim 1 wherein each recognition
means recognizes a set of target subsequences, and wherein a
sequence from said database matches a generated signal when
the sequence from said database has both the same length
10 between occurrences of target subsequences as is represented
by the generated signal, and the target subsequences are
members of the sets of target subsequences represented by the
generated signal.

15   4.   The method of claim 1 further comprising dividing
said sample of nucleic acids into a plurality of portions and
performing the steps of claim 1 individually on a plurality
of said portions, wherein a different one or more recognition
means are used with each portion.
20

        5.   The method of claim 1 wherein the quantitative
abundance of nucleic acids containing said nucleotide
sequences in the sample is determined from the quantitative
level of the one or more signals determined to match said
25 sequences.

        6.   The method of claim 1 wherein said plurality of
nucleic acids are DNA.

30   7.   The method of claim 6 wherein the DNA is cDNA.

        8.   The method of claim 7 wherein the cDNA is prepared
from a plant, a single celled animal, a multicellular animal,
a bacterium, a virus, a fungus, or a yeast.
35

        9.   The method of claim 8 wherein said database
comprises substantially all the known xpr ss d s quenc s of

said plant, single celled animal, multicellular animal,
bacterium, virus, fungus, or yeast.

10.   The method of claim 7 wherein the cDNA is of total
5 cellular RNA or total cellular poly(A) RNA.

11.   The method of claim 6 wherein the recognition
means are one or more restriction endonucleases whose
recognition sites are said target subsequences, and wherein
10   the step of probing comprises digesting said sample
with said one or more restriction endonucleases into
fragments and ligating double stranded adapter DNA molecules
to said fragments to produce ligated fragments, each said
adapter DNA molecule comprising (i) a shorter stand having no
15 5' terminal phosphates and consisting of a first and second
portion, said first portion at the 5' end of the shorter
strand and being complementary to the overhang produced by
one of said restriction endonucleases, and (ii) a longer
strand having a 3' end subsequence complementary to said
20 second portion of the shorter strand; and wherein
        the step of generating further comprises melting the
shorter strand from the ligated fragments, contacting the
ligated fragments with a DNA polymerase, extending the
ligated fragments by synthesis with the DNA polymerase to
25 produce blunt-ended double stranded DNA fragments, and
amplifying the blunt-ended fragments by a method comprising
contacting the blunt-ended fragments with the DNA polymerase
and primer oligodeoxynucleotides, said primer
oligodeoxynucleotides comprising the longer adapter strand,
30 and said contacting being at a temperature not greater than
the melting temperature of the primer oligodeoxynucleotide
from a strand of the blunt-ended fragments complementary to
the primer oligodeoxynucleotide and not less than the melting
t mperatur of the shorter strand of the adapt r nucleic acid
35 from the blunt-ended fragments.

12.   The method of claim 6 wherein the recognition
means are one or more restriction endonucleases whose
recognition sites are said target subsequences, and wherein
the step of probing further comprises digesting the sample
5 with said one or more restriction endonucleases.

13.   The method of claim 12 further comprising:
(a)   identifying a fragment of a nucleic acid in the
sample which generates said one or more signals; and
10   (b)   recovering said fragment.

14.   The method of claim 13 wherein the signals
generated by said recovered fragment do not match a sequence
in said nucleotide sequence database.
15

15.   The method of claim 13 which further comprises
using at least a hybridizable portion of said fragment as a
hybridization probe to bind to a nucleic acid that can
generate said fragment upon digestion by said one or more
20 restriction endonucleases.

16.   The method of claim 12 wherein the step of
generating further comprises after said digesting: removing
from the sample both nucleic acids which have not been
25 digested and nucleic acid fragments resulting from digestion
at only a single terminus of the fragments.

17.   The method of claim 16 wherein prior to digesting,
the nucleic acids in the sample are each bound at one
30 terminus to a biotin molecule, and said removing is carried
out by a method which comprises contacting the nucleic acids
in the sample with streptavidin or avidin affixed to a solid
support.

35   18.   The m thod of claim 16 wh rein prior to digesting,
th nucleic acids in the sampl are each bound at ne
t rminus to a hapt n molecul , and said removing is carri d

out by a method which comprises contacting the nucleic acids in the sampl with an anti-hapt n antibody affixed to a solid support.

5   19. The method of claim 12 wherein said digesting with said one or more restriction endonucleases leaves single-stranded nucleotide overhangs on the digested ends.

           20. The method of claim 19 wherein the step of probing
10 further comprises hybridizing double-stranded adapter nucleic acids with the digested sample fragments, each said adapter nucleic acid having an end complementary to said overhang generated by a particular one of the one or more restriction endonucleases, and ligating with a ligase a strand of said
15 adapter nucleic acids to the 5' end of a strand of the digested sample fragments to form ligated nucleic acid fragments.

           21. The method of claim 20 wherein said digesting with
20 said one or more restriction endonucleases and said ligating are carried out in the same reaction medium.

           22. The method of claim 21 wherein said digesting and said ligating comprises incubating said reaction medium at a
25 first temperature and then at a second temperature, wherein said one or more restriction endonucleases are more active at the first temperature than the second temperature and said ligase is more active at the second temperature than the first temperature.
30
           23. The method of claim 22 wherein said incubating at said first temperature and said incubating at said second temperature are performed repetitively.

35   24. The method f claim 20 wh rein th st p of probing further compris s prior to said digesting: r moving terminal

phosphates from DNA in said sample by incubation with an
alkaline phosphatas .

25. The method of claim 24 wherein said alkaline
5 phosphatase is heat labile and is heat inactivated prior to
said digesting.

26. The method of claim 20 wherein said generating
step comprises amplifying the ligated nucleic acid fragments.
10

27. The method of claim 26 wherein said amplifying is
carried out by use of a nucleic acid polymerase and primer
nucleic acid strands, said primer nucleic acid strands being
capable of priming nucleic acid synthesis by said polymerase.
15

28. The method of claim 27 wherein the primer nucleic
acid strands have a G+C content of between 40% and 60%.

29. The method of claim 27 wherein each said adapter
20 nucleic acid has a shorter strand and a longer strand, the
longer strand being ligated to the digested sample fragments,
and said generating step comprises prior to said amplifying
step the melting of the shorter strand from the ligated
fragments, contacting the ligated fragments with a DNA
25 polymerase, extending the ligated fragments by synthesis with
the DNA polymerase to produce blunt-ended double stranded DNA
fragments, and wherein the primer nucleic acid strands
comprise a hybridizable portion of the sequence of said
longer strands, each different primer nucleic acid strand
30 priming amplification only of blunt ended double stranded DNA
fragments that are produced after digestion by a particular
restriction endonuclease.

30. The method of claim 27 wherein each said adapter
35 nucleic acid has a shorter strand and a longer strand, the
long r strand being ligated to the dig sted sample fragm nts,
and said generating st p comprises prior to said amplifying

- 260 -

step the melting of the shorter strand from the ligat d
fragments, contacting the ligat d fragments with a DNA
polymerase, extending the ligated fragments by synthesis with
the DNA polymerase to produce blunt-ended double stranded DNA
5 fragments, and wherein the primer nucleic acid strands
comprise the sequence of said longer strands, each different
primer nucleic acid strand priming amplification only of
blunt ended double stranded DNA fragments that are produced
after digestion by a particular restriction endonuclease.
10

31.   The method of claim 30 wherein during said
amplifying step the primer nucleic acid strands are annealed
to the ligated nucleic acid fragments at a temperature that
is less than the melting temperature of the primer nucleic
15 acid strands from strands complementary to the primer nucleic
acid strands but greater than the melting temperature of the
shorter adapter strands from said blunt-ended fragments.

32.   The method of claim 30 wherein the primer nucleic
20 acid strands comprise primers, each primer specific for a
particular restriction endonuclease, and further comprising
at the 3' end of and contiguous with the longer strand
sequence, the portion of the restriction endonuclease
recognition site remaining on a nucleic acid fragment
25 terminus after digestion by the restriction endonuclease.

33.   The method of claim 32 wherein each said primer
specific for a particular restriction endonuclease further
comprises at its 3' end one or more nucleotides 3' to and
30 contiguous with the remaining portion of the restriction
endonuclease recognition site, whereby the ligated nucleic
acid fragment amplified is that comprising said remaining
portion of said restriction endonuclease recognition site
contiguous to said on or more additional nucleotides.
35

34.   The method f claim 33 wherein said specific
primers are detectably labeled, such that said primers

comprising a particular said one or more additional
nucleotides can be distinguishably detected from said primers
comprising a different said one or more additional
nucleotides.

35.   The method of claim 6 wherein the recognition
means are oligomers of nucleotides, nucleotide-mimics, or a
combination of nucleotides and nucleotide-mimics, which are
specifically hybridizable with the target subsequences.

36.   The method of claim 35 wherein the step of
generating comprises amplifying with a nucleic acid
polymerase and with primers comprising said oligomers,
whereby fragments of nucleic acids in the sample between
hybridized oligomers are amplified.

37.   The method of claim 36 further comprising:
(a)   identifying a fragment of a nucleic acid in the
sample which generates said one or more signals; and
(b)   recovering said fragment.

38.   The method of claim 37 wherein the signals
generated by said recovered fragment do not match a sequence
in said nucleotide database.

39.   The method of claim 37 which further comprises
using at least a hybridizable portion of said fragment as a
hybridization probe to bind to a nucleic acid that can
generate said fragment upon amplification with said nucleic
acid polymerase and said one or more primers.

40.   The method of claim 1 wherein said signals further
comprise a representation of whether an additional target
subsequence is present on said nucleic acid in the sample
betw n said occurrences of targ t subsequ nces.

41. The method of claim 40 wherein said additional
target subsequence is recognized by a method comprising
contacting nucleic acids in the sample with oligomers of
nucleotides, nucleotide-mimics, or mixed nucleotides and
5 nucleotide-mimics, which are hybridizable with said
additional target subsequence.

42. The method of claim 1 wherein the step of
generating comprises suppressing said signals when an
10 additional target subsequence is present on said nucleic acid
in the sample between said occurrences of target
subsequences.

43. The method of claim 42 wherein the step of
15 generating comprises amplifying nucleic acids in the sample,
and wherein said additional target subsequence is recognized
by a method comprising contacting nucleic acids in the sample
with (a) oligomers of nucleotides, nucleotide-mimics, or
mixed nucleotides and nucleotide-mimics, which hybridize with
20 said additional target subsequence and disrupt the amplifying
step; or (b) restriction endonucleases which have said
additional target subsequence as a recognition site and
digest the nucleic acids in the sample at the recognition
site.
25

44. The method of claim 12 or 36 wherein the step of
generating further comprises separating nucleic acid
fragments by length.

30   45. The method of claim 44 wherein the step of
generating further comprises detecting said separated nucleic
acid fragments.

46. The method of claim 45 wherein the quantitative
35 abundance of a nucleic acid comprising a particular
nucleotide sequence in the sample is determined from the
quantitative level of the one or more signals gen rat d by

said nucleic acid that are determined to match said
particular nucleotide sequence.

47.   The method of claim 45 wherein said detecting is
carried out by a method comprising staining said fragments
with silver, labeling said fragments with a DNA intercalating
dye, or detecting light emission from a fluorochrome label on
said fragments.

48.   The method of claim 45 wherein said representation
of the length between occurrences of target subsequences is
the length of fragments determined by said separating and
detecting steps.

49.   The method of claim 45 wherein said separating is
carried out by use of liquid chromatography or mass
spectrometry.

50.   The method of claim 45 wherein said separating is
carried out by use of electrophoresis.

51.   The method of claim 50 wherein said
electrophoresis is carried out in a slab gel or capillary
configuration using a denaturing or non-denaturing medium.

52.   The method of claim 1 wherein a predetermined one
or more nucleotide sequences in said database are of
interest, and wherein the target subsequences are such that
said sequences of interest generate at least one signal that
is not generated by other nucleotide sequences in said
database.

53.   The method of claim 52 wherein the nucleotide
sequences of interest are a majority of the sequences in said
database.

54. The method of claim 1 wher in the target subsequences have a probability of occurrence in the nucleotide sequences in said database of from approximately 0.01 to approximately 0.30.

55. The method of claim 1 wherein the target subsequences are such that nucleotide sequences in said database contain on average a sufficient number of occurrences of target subsequences in order to on average generate a signal that is not generated by any other nucleotide sequence in said database.

56. The method of claim 55 wherein the number of pairs of target subsequences present on average in a nucleotide sequence in said database is no less than 3, and wherein the average number of signals generated from nucleotide sequences in said database is such that the average difference between lengths represented by the generated signals is greater than or equal to 1 nucleotide.

57. The method of claim 55 wherein the target subsequences have a probability of occurrence, p, approximately given by the solution of

$$\frac{R(R + 1)p^2}{2} = A$$

and

$$\frac{L}{Np^2} = B$$

wherein N = the number of different nucleotide sequences in said database; L = the average length of said different nucleotide sequences in said database; R = the number of r cognition means; A = the number of pairs of target subsequences present on average in said different nucleotid sequences in said databas ; and B = th average difference

between lengths r pr sented by the signals generated from the s quenc s in said database.

58.   The method of claim 57 wherein A is greater than or equal to 3.

59.   The method of claim 57 wherein B is greater than or equal to 1.

60.   The method of claim 1 wherein the target subsequences are selected according to the further steps comprising:

(a)   determining a pattern of signals that can be generated and the sequences capable of generating each such signal by simulating the steps of probing and generating applied to sequences in said database of nucleotide sequences;

(b)   ascertaining the value of said determined pattern according to an information measure; and

(c)   choosing the target subsequences in order to generate a new pattern that optimizes the information measure.

61.   The method of claim 60 wherein said choosing step selects target subsequences which comprise the recognition sites of the one or more restriction endonucleases.

62.   The method of claim 60 wherein said choosing step selects target subsequences which comprise the recognition sites of the one or more restriction endonucleases contiguous with one or more additional nucleotides.

63.   The method of claim 60 wherein a predetermined one or more of the nucleotide sequences present in said database of nucl otide sequenc s ar of int r st, and the information measure optimized is the numb r of such said sequences of interest which generate at l ast one signal that is not

generated by any other nucleotide sequence present in said database.

64.   The method of claim 63 wherein said nucleotide
5 sequences of interest are a majority of the nucleotide sequences present in said database.

65.   The method of claim 60 wherein said choosing step is by exhaustive search of all combinations of target
10 subsequences of length less than approximately 10.

66.   The method of claim 60 wherein said step of choosing target subsequences is by a method comprising simulated annealing.
15
67.   The method of claim 1 wherein the step of searching further comprises:
      (a)   determining a pattern of signals that can be generated and the sequences capable of generating each such
20 signal by simulating the steps of probing and generating applied to each sequence in said database of nucleotide sequences; and
      (b)   finding the one or more nucleotide sequences in said database that are able to generate said one or more
25 generated signals by finding in said pattern those signals that comprise a representation of (i) the same lengths between occurrences of target subsequences as is represented by the generated signal, and (ii) the same target subsequences as are represented by the generated signal, or
30 target subsequences that are members of the same sets of target subsequences represented by the generated signal.

68.   The method of claim 60 or 67 wherein the step of determining further comprises:
35   (a)   s arching for occurr nc s of said target subsequences or sets of target subs quenc s in nucleotide sequenc s in said database of nucleotide s qu nc s;

(b) finding the l ngths between occurrenc s of said
target subsequences or sets of target subsequences in the
nucleotide sequences of said database; and

(c) forming the pattern of signals that can be
5 generated from the sequences of said database in which the
target subsequences were found to occur.

69. The method of claim 20 wherein said restriction
endonucleases generate 5' overhangs at the terminus of
10 digested fragments and wherein each double stranded adapter
nucleic acid comprises:

(a) a shorter nucleic acid strand consisting of a
first and second contiguous portion, said first portion being
a 5' end subsequence complementary to the overhang produced
15 by one of said restriction endonucleases; and

(b) a longer nucleic acid strand having a 3' end
subsequence complementary to said second portion of the
shorter strand.

20   70. The method of claim 69 wherein said shorter strand
has a melting temperature from a complementary strand of less
than approximately 68 °C, and has no terminal phosphate.

71. The method of claim 70 wherein said shorter strand
25 is approximately 12 nucleotides long.

72. The method of claim 69 wherein said longer strand
has a melting temperature from a complementary strand of
greater than approximately 68 °C, is not complementary to any
30 nucleotide sequence in said database, and has no terminal
phosphate.

73. The method of claim 72 wherein said ligated
nucleic acid fragments do not contain a recognition site for
35 any of said restriction endonucleases.

74.   Th   method of claim 72 wherein said one or more restriction endonucleases are heat inactivated before said ligating.

5   75.   The method of claim 72 wherein said longer strand is approximately 24 nucleotides long and has a G+C content between 40% and 60%.

          76.   The method of claim 20 wherein said restriction
10 endonucleases generate 3' overhangs at the terminus of the digested fragments, and wherein each double stranded adapter nucleic acid comprises:
          (a)   a longer nucleic acid strand consisting of a first and second contiguous portion, said first portion being a 3'
15 end subsequence complementary to the overhang produced by one of said restriction endonucleases; and
          (b)   a shorter nucleic acid strand complementary to the 3' end of said second portion of the longer nucleic acid stand.
20

          77.   The method of claim 76 wherein said shorter strand has a melting temperature from said longer strand of less than approximately 68 °C, and has no terminal phosphates.

25   78.   The method of claim 77 wherein said shorter strand is 12 base pairs long.

          79.   The method of claim 76 wherein said longer strand has a melting temperature from a complementary strand of
30 greater than approximately 68 °C, is not complementary to any nucleotide sequence in said database, has no terminal phosphate, and wherein said ligated nucleic acid fragments do not contain a recognition site for any of said restriction endonucleases.
35

80. The method of claim 79 wher in said longer strand
is 24 base pairs long and has a G+C content between 40% and
60%.

5   81. A method for identifying or classifying a nucleic
acid comprising:
         (a) probing said nucleic acid with a plurality of
recognition means, each recognition means recognizing a
target nucleotide subsequence or a set of target nucleotide
10 subsequences, in order to generate a set of signals, each
signal representing whether said target subsequence or one of
said set of target subsequences is present or absent in said
nucleic acid; and
         (b) searching a nucleotide sequence database, said
15 database comprising a plurality of known nucleotide sequences
of nucleic acids that may be present in the sample, for
sequences matching said generated set of signals, a sequence
from said database matching a set of signals when the
sequence from said database (i) comprises the same target
20 subsequences as are represented as present, or comprises
target subsequences that are members of the sets of target
subsequences represented as present by the generated sets of
signals, and (ii) does not comprise the target subsequences
represented as absent or that are members of the sets of
25 target subsequences represented as absent by the generated
sets of signals,
         whereby the nucleic acid is identified or classified.

         82. The method of claim 81 wherein the set of signals
30 are represented by a hash code which is a binary number.

         83. The method of claim 81 wherein the step of probing
generates quantitative signals of the numbers of occurrences
of said target subsequences or of members of said set of
35 targ t subs quences in said nucleic acid.

84. The method of claim 83 wh rein a sequ nce matches said generated set of signals when the sequence from said database comprises the same target subsequences with the same number of occurrences in said sequence as in the quantitative

5 signals and does not comprise the target subsequences represented as absent or target subsequences within the sets of target subsequences represented as absent.

85. The method of claim 81 wherein said plurality of

10 nucleic acids are DNA.

86. The method of claim 85 wherein the recognition means are detectably labeled oligomers of nucleotides, nucleotide-mimics, or combinations of nucleotides and ·

15 nucleotide-mimics, and the step of probing comprises hybridizing said nucleic acid with said oligomers.

87. The method of claim 86 wherein said detectably labeled oligomers are detected by a method comprising

20 detecting light emission from a fluorochrome label on said oligomers, or arranging said labeled oligomers to cause light to scatter from a light pipe and detecting said scattering.

88. The method of claim 86 wherein the recognition

25 means are oligomers of peptido-nucleic acids.

89. The method of claim 86 wherein the recognition means are DNA oligomers, DNA oligomers comprising universal nucleotides, or sets of partially degenerate DNA oligomers.

30

90. The method of claim 85 wherein the step of searching further comprises:
(a) determining a pattern of sets of signals of the presence or absence of said target subsequences or said sets

35 of target subsequ nc s that can be generated and the sequenc s capable of gen rating each set of signals in said

pattern by simulating the step of probing as applied to each
s quenc in said database of nucleotide sequ nces; and

      (b)   finding one or more nucleotide sequences that are
capable of generating said generated set of signals by
5 finding in said pattern those sets that match said generated
set, where a set of signals from said pattern matches a
generated set of signals when the set from said pattern (i)
represents as present the same target subsequences as are
represented as present or target subsequences that are
10 members of the sets of target subsequences represented as
present by the generated sets of signals and (ii) represents
as absent the target subsequences represented as absent or
that are members of the sets of target subsequences
represented as absent by the generated sets of signals.
15

      91.   The method of claim 85 wherein the target
subsequences are selected according to the further steps
comprising:

      (a)   determining (i) a pattern of sets of signals
20 representing the presence or absence of said target
subsequences or of said sets of target subsequences that can
be generated, and (ii) the sequences capable of generating
each set of signals in said pattern by simulating the step of
probing as applied to each sequence in said database of
25 nucleotide sequences;

      (b)   ascertaining the value of said pattern generated
according to an information measure; and

      (c)   choosing the target subsequences in order to
generate a new pattern that optimizes the information
30 measure.

      92.   The method of claim 91 wherein the information
measure is the number of sets of signals in the pattern which
are capable of being generated by one or more sequences in
35 said database.

93. The method of claim 91 wher in the information
measure is the number of s ts of signals in the pattern which
are capable of being generated by only one sequence in said
database.

5

94. The method of claim 91 wherein said choosing step
is by a method comprising exhaustive search of all
combination of target subsequences of length less than
approximately 10.

10

95. The method of claim 91 wherein said choosing step
is by a method comprising simulated annealing.

96. The method of claim 90 or 91 wherein the step of
15 determining by simulating further comprises:
(a) searching for the presence or absence of said
target subsequences or sets of target subsequences in each
nucleotide sequence in said database of nucleotide sequences;
and
20   (b) forming the pattern of sets of signals that can be
generated from said sequences in said database.

97. The method of claim 96 where the step of searching
is carried out by a string search.

25

98. The method of claim 96 wherein the step of
searching comprises counting the number of occurrences of
said target subsequences in each nucleotide sequence.

30   99. The method of claim 81 wherein the target
subsequences have a probability of occurrence in a nucleotide
sequence in said database of nucleotide sequences of from
0.01 to 0.6.

35   100. The meth d of claim 99 wher in the target
subsequences are such that the presence of one target
subsequence in a nucleotid sequ nc in said database of

- 273 -

nucleotide sequences is substantially independent of the
presence of any other target subsequence in the nucleotide
sequence.

5   101.   The method of claim 99 wherein fewer than
approximately 50 target subsequences are selected.

         102.   A programmable apparatus for analyzing signals
comprising:
10   (a)   an inputting device for inputting one or more
actual signals generated by probing a sample comprising a
plurality of nucleic acids with recognition means, each
recognition means recognizing a target nucleotide subsequence
or a set of target nucleotide subsequences, said signals
15 comprising a representation of (i) the length between
occurrences of said target subsequences in a nucleic acid of
said sample, and (ii) the identities of said target
subsequences in said nucleic acid, or the identities of said
sets of target subsequences among which is included the
20 target subsequences in said nucleic acid;
         (b)   a searching device operatively coupled to said
accepting device for searching a sequence in a nucleotide
sequence database for occurrences of said target subsequences
or target subsequences that are members of said sets of
25 target subsequences, and for the length between such
occurrences, said database comprising a plurality of known
nucleotide sequences that may be present in said sample;
         (c)   a comparing device operatively coupled to said
accepting device and to said searching device for finding a
30 match between said one or more actual signals and a sequence
in said database, said one or more actual signals matching a
sequence from said database when the sequence from said
database has both (i) the same length between occurrences of
target subsequences as is represented by said one or more
35 actual signals, and (ii) the same targ t subsequences as are
repr sented by said on or mor actual signals, or target
subs quences that are members of the sets of target

- 274 -

subsequences represented by said one or more actual signals; and

(d)   a control device operatively coupled to said comparing device for causing said comparing to be done for sequences in the database and for outputting those database sequences that match said one or more actual signals.

103.   The programmable apparatus of claim 102 wherein said searching device searches for said target subsequences or a set of target nucleotide subsequences in said database sequences by performing a string comparison of the nucleotides in said subsequences with those in said database sequence.

104.   The programmable apparatus of claim 102 wherein said control device further comprises causing said searching device to search all sequences in said database in order to determine a pattern of signals that can be generated by probing said sample with said recognition means, and wherein said control device further causes said comparing device to find any matches between said one or more actual signals and said pattern of signals, said one or more actual signals matching a signal in said pattern of signals when the signal from said pattern represents (i) the same length between occurrences of target subsequences as is represented by said one or more actual signals, and (ii) the same target subsequences as are represented by said one or more actual signals, or target subsequences that are members of the sets of target subsequences represented by said one or more actual signals.

105.   The programmable apparatus of claim 102 wherein said sample of nucleic acids comprises cDNA of RNA of a cell or tissue type, and said database comprises DNA sequences that are likely to b xpressed by said cell or tissue type.

106. A computer readable memory that can be used to
direct a programmable apparatus to function for analyzing
signals according to steps comprising:

(a) inputting one or more actual signals generated by
5 probing a sample comprising a plurality of nucleic acids with
recognition means, each recognition means recognizing a
target nucleotide subsequence or a set of target nucleotide
subsequences, said signals comprising a representation of (i)
the length between occurrences of said target subsequences in
10 a nucleic acid of said sample, and (ii) the identities of
said target subsequences in said nucleic acid, or the
identities of said sets of target subsequences among which is
included the target subsequences in said nucleic acid;

(b) searching a sequence in a nucleotide sequence
15 database for occurrences of said target subsequences or
target subsequences that are members of said sets of target
subsequences, and for the length between such occurrences,
said database comprising a plurality of known nucleotide
sequences that may be present in said sample;

20   (c) matching said one or more actual signals and a
sequence in said database when the sequence in said database
has both (i) the same length between occurrences of target
subsequences as is represented by said one or more actual
signals and (ii) the same target subsequences as are
25 represented by said one or more actual signals, or target
subsequences that are members of the sets of target
subsequences as are represented by said one or more actual
signals; and

(d) repetitively performing said searching and matching
30 steps for the majority of sequences in the database and
outputting those database sequences that match said one or
more actual signals.

107. A programmable apparatus for selecting target
35 subsequences comprising:

(a) an initial s lection devic for selecting initial
target subsequences or initial s ts of target subs quenc s;

(b)   a first control device;

(c)   a search device operatively coupled to said initial
selection device and to said first control device (i) for
searching sequences in a nucleotide sequence database for
5 occurrences of said initial target subsequences or
occurrences of target subsequences that are members of said
initial sets of target subsequences and for the length
between such occurrences, and (ii) for determining an initial
pattern of signals that can be generated from said selected
10 initial target subsequences or said initial sets of target
subsequences, said database comprising a plurality of known
nucleotide sequences, said signals comprising a
representation of (i) the length between said occurrences in
a sequence in said database, and (ii) the identities of said
15 initial target subsequences that occur in said sequence in
said database, or the identities of target subsequences that
are members of the initial sets of target subsequences that
occur in said sequence in said database; and

(d)   an ascertaining device operatively coupled to said
20 searching device and to said first control device for
ascertaining the value of said determined initial pattern
according to an information measure; and wherein

said first control device causes further target
subsequences to be selected and causes the search device to
25 determine a further pattern of signals and the ascertaining
device to ascertain a further value of said information
measure and accepts the further target subsequences when said
further pattern optimizes said further value of said
information measure.

30

108.   The programmable apparatus of claim 107 wherein a
predetermined one or more of the sequences in said database
are of interest, and wherein said ascertaining device
ascertains the value of an information measure by counting
35 the number of such s quences of interest which generate in
said det rmin d pattern at least one signal that is not
generated by any oth r sequence in said database.

109.   The programmable apparatus of claim 108 wherein
said one or more of the sequences of interest comprise
substantially all the sequences in said database.

5   110.   The programmable apparatus of claim 107 wherein
said first control device optimizes the value of said
information measure according to a method of exhaustive
search, wherein said first control device selects further
target subsequences of length less than approximately 10 and
10 accepts the further target subsequences if said further value
of said information measure is greater than the previous
value.

111.   The programmable apparatus of claim 107 wherein
15 said first control device optimizes the value of said
information measure according to a method comprising
simulated annealing, wherein said first control device
repeatedly selects further target subsequences and accepts
the further target subsequences if said further value of said
20 information measure is not decreased by greater than a
probabilistic factor dependent on a simulated-temperature,
and wherein said programmable apparatus further comprises a
second control device operatively coupled to said first
control device for decreasing said simulated-temperature as
25 said first control device selects further target
subsequences.

112.   The programmable apparatus of claim 111 wherein
said probabilistic factor is an exponential function of the
30 negative of the decrease in the information measure divided
by said simulated-temperature.

113.   The programmable apparatus of claim 107 wherein
said database comprises a majority of known DNA sequences
35 that are likely to be expressed in one or more cell types.

- 278 -

114. A computer readable memory that can be us d to
direct a programmable apparatus to function for s l cting
target subsequences according to steps comprising:

(a)   selecting initial target subsequences or initial
5 sets of target subsequences;

(b)   searching a sequence in a nucleotide sequence
database for occurrences of said initial target subsequences
or occurrences of target subsequences that are members of
said initial sets of target subsequences and for the length
10 between such occurrences, said database comprising a
plurality of known nucleotide sequences that may be present
in said sample;

(c)   .determining an initial pattern of signals that can
be generated from said selected initial target subsequences
15 or said initial sets of target subsequences, said signals
comprising a representation of (i) the length between said
occurrences in a sequence in said database, and (ii) the
identities of said initial target subsequences that occur in
said sequence in said database, or the identities of target
20 subsequences that are members of the initial sets of target
subsequences that occur in said sequence in said database;
and

(d)   ascertaining the value of said determined initial
pattern according to an information measure; and

25   (e)   repetitively performing said selecting, searching,
determining, and ascertaining steps to determine a further
pattern of signals and a further value of said information
measure, and accepting the further target subsequences when
said further pattern optimizes said further value of said
30 information measure.

115. A programmable apparatus for displaying data
comprising:

(a)   a selecting device for selecting target
35 subs quences or sets of target subsequences, such that
recognition m ans for recognizing said target subsequ nces or
said sets of target subsequences can be used to generate

signals by probing a sample comprising a plurality of nucleic
acids, said signals comprising a repres ntation of (i) the
length between occurrences of said target subsequences in a
nucleic acid of said sample and (ii) the identities of said
5 target subsequences in said nucleic acid or the identities of
said sets of target subsequences among which are included the
target subsequences in said nucleic acid;

      (b)   an inputting device for inputting one or more
actual signals generated by probing said sample with said
10 recognition means;

      (c)   an analyzing device for analyzing signals
operatively coupled to said selecting and inputting devices
that determines which sequences in a nucleotide sequence
database can generate said actual signals when subject to
15 said recognition means, said database comprising a plurality
of known nucleotide sequences that may be present in said
sample;

      (d)   an input/output device operatively coupled to said
selecting, inputting, and analyzing devices that inputs user
20 requests and controls the selecting device to select target
subsequences or sets of target subsequences, controls the
inputting device to accept actual signals, controls the
analyzing device to find the sequences in said database that
can generate said actual signals, and displays output
25 comprising said actual signals and said sequences in said
database that can generate said actual signals.

      116.   The programmable apparatus of 115 wherein said
sample is a cDNA sample prepared from a tissue specimen, and
30 the apparatus further comprises a storage device operatively
coupled to the input/output device for storing indications of
the origin of said tissue specimen and information concerning
said tissue specimen,

      and wherein said indications can be displayed upon user
35 input.

117. The programmable apparatus of 116 wherein the indications and information concerning said tissue specimen comprises histological information comprising tissue images.

5   118. The programmable apparatus of claim 115 further comprising:

(a) one or more instrument devices for probing said sample with said recognition means and for generating said actual signals; and

10   (b) a control device operatively coupled to said one or more instrument devices and to said input/output device for controlling the operation of said instrument devices,

wherein said user can input control commands for control of said instrument devices and receive output concerning the

15 status of said instrument devices.

119. The programmable apparatus of 118 wherein the one or more instrument devices are capable of automatic operation, whereby the probing and generating can be

20 performed without manual intervention.

120. The programmable apparatus of claim 115 wherein one or more of said selecting, inputting, analyzing, and input/output devices are physically collocated with each

25 other.

121. The programmable apparatus of claim 115 wherein one or more of said selecting, inputting, analyzing, and input/output devices are physically spaced apart from each

30 other and are connected by a communication medium for exchanges of commands and information.

122. A computer readable memory that can be used to direct a programmabl apparatus to function for displaying

35 data according to steps comprising:

(a) selecting target subsequences or sets of target subsequences, such that r cognition means for recognizing

said targ t subsequences or said sets of target subsequ nces
can be used to generate signals by probing a sample
comprising a plurality of nucleic acids, said signals
comprising a representation of (i) the length between
5 occurrences of said target subsequences in a nucleic acid of
said sample and (ii) the identities of said target
subsequences in said nucleic acid or the identities of said
sets of target subsequences among which are included the
target subsequences in said nucleic acid;

10   (b)   inputting one or more actual signals generated by
probing said sample with said recognition means;

        (c)   analyzing said one or more actual signals to
determine which sequences in a nucleotide sequence database
can generate said actual signals when subject to said
15 recognition means, said database comprising a plurality of
known nucleotide sequences that may be present in said
sample; and

        (d)   inputting user requests to control said selecting
step to select target subsequences or sets of target
20 subsequences, said inputting step to input actual signals,
and said analyzing step to find the sequences in said
database that can generate said actual signals, and
outputting in response to further user requests information
comprising said actual signals and said sequences in said
25 database that can generate said actual signals.

        123. A method for identifying, classifying, or
quantifying DNA molecules in a sample of DNA molecules having
a plurality of different nucleotide sequences, the method
30 comprising the steps of:

        (a)   digesting said sample with one or more restriction
endonucleases, each said restriction endonuclease recognizing
a subsequence recognition site and digesting DNA at said
recognition site to produce fragments with 5' overhangs;
35   (b)   contacting said fragm nts with short r and longer
oligode xynucleotides, each said short r ligodeoxynucleotide
hybridizable with a said 5' overhang and having no t rminal

- 282 -

phosphates, each said longer oligodeoxynucleotide
hybridizable with a said shorter oligodeoxynucleotide;

     (c) ligating said longer oligodeoxynucleotides to said
5' overhangs on said DNA fragments to produce ligated DNA
5 fragments;

     (d) extending said ligated DNA fragments by synthesis
with a DNA polymerase to produce blunt-ended double stranded
DNA fragments;

     (e) amplifying said blunt-ended double stranded DNA
10 fragments by a method comprising contacting said DNA
fragments with a DNA polymerase and primer
oligodeoxynucleotides, each said primer oligodeoxynucleotide
having a sequence comprising that of one of the longer
oligodeoxynucleotides;

15   (f) determining the length of the amplified DNA
fragments; and

     (g) searching a DNA sequence database, said database
comprising a plurality of known DNA sequences that may be
present in the sample, for sequences matching one or more of
20 said fragments of determined length, a sequence from said
database matching a fragment of determined length when the
sequence from said database comprises recognition sites of
said one or more restriction endonucleases spaced apart by
the determined length,

25   whereby DNA molecules in said sample are identified,
classified, or quantified.

     124. The method of claim 123 wherein the sequence of
each primer oligodeoxynucleotide further comprises 3' to and
30 contiguous with the sequence of the longer
oligodeoxynucleotide the portion of the recognition site of
said one or more restriction endonucleases remaining on a DNA
fragment terminus after digestion, said remaining portion
being 5' to and contiguous with one or more additional
35 nucleotides, and wher in a sequence from said database
matches a fragment of det rmined length when the s qu nce
from said database comprises subsequences that are the

recognition sites of said one or more restriction
endonucleases contiguous with said one or more additional
nucleotides and when the subsequences are spaced apart by the
determined length.

5

125. The method of claim 123 wherein said determining
step further comprises detecting the amplified DNA fragments
by a method comprising staining said fragments with silver.

10   126. The method of claim 123 wherein said longer
oligodeoxynucleotides are detectably labeled, wherein the
determining step further comprises detection of said
detectable labels, and wherein a sequence from said database
matches a fragment of determined length when the sequence
15 from said database comprises recognition sites of the one or
more restriction endonucleases, said recognition sites being
identified by the detectable labels of said longer
oligodeoxynucleotides, said recognition sites being spaced
apart by the determined length.
20

127. The method of claim 123 wherein said determining
step further comprises detecting the amplified DNA fragments
by a method comprising labeling said fragments with a DNA
intercalating dye or detecting light emission from a
25 fluorochrome label on said fragments.

128. The method of claim 123 further comprising, prior
to said determining step, the step of hybridizing the
amplified DNA fragments with a detectably labeled
30 oligodeoxynucleotide complementary to a subsequence, said
subsequence differing from said recognition sites of said one
or more restriction endonucleases, wherein the determining
step further comprises detecting said detectable label of
said oligodeoxynucleotide, and wherein a sequence from said
35 database matches a fragment of determined length when the
sequence from said database furth r comprises said

subsequence between the recognition sites of said on or more
restriction endonucleases.

129.   The method of claim 123 wherein the one or more
restriction endonucleases are pairs of restriction
endonucleases, the pairs being selected from the group
consisting of Acc56I and HindIII, Acc65I and NgoMI, BamHI and
EcoRI, BglII and HindIII, BglII and NgoMI, BsiWI and BspHI,
BspHI and BstYI, BspHI and NgoMI, BsrGI and EcoRI, EagI and
EcoRI, EagI and HindIII, EagI and NcoI, HindIII and NgoMI,
NgoMI and NheI, NgoMI and SpeI, BglII and BspHI, Bsp120I and
NcoI, BssHII and NgoMI, EcoRI and HindIII, and NgoMI and
XbaI.

130.   The method of claim 123 wherein the step of
ligating is performed with T4 DNA ligase.

131.   The method of claim 123 wherein the steps of
digesting, contacting, and ligating are performed
simultaneously in the same reaction vessel.

132.   The method of claim 123 wherein the steps of
digesting, contacting, ligating, extending, and amplifying
are performed in the same reaction vessel.

133.   The method of claim 123 wherein the step of
determining the length is performed by electrophoresis.

134.   The method of claim 123 wherein the step of
searching said DNA database further comprises:
    (a)   determining a pattern of fragments that can be
generated and for each fragment in said pattern those
sequences in said DNA database that are capable of generating
the fragm nt by simulating th steps of digesting with said
one or mor restriction endonucleases, contacting, ligating,
extending, amplifying, and determining applied to each
sequence in said DNA databas ; and

- 285 -

(b) finding the sequences that are capable of generating said one or more fragments of determined length by finding in said pattern one or more fragments that have the same length and recognition sites as said one or more
5 fragments of determined length.

135. The method of claim 123 wherein the steps of digesting and ligating go substantially to completion.

10   136. The method of claim 123 wherein the DNA sample is cDNA of RNA from a tissue or a cell type derived from a plant, a single celled animal, a multicellular animal, a bacterium, a virus, a fungus, or a yeast.

15   137. The method of claim 123 wherein the DNA sample is cDNA of RNA from one or more cell types of a mammal.

138. The method of claim 137 wherein the mammal is a human.
20

139. The method of claim 137 wherein the mammal is a human having or suspected of having a diseased condition.

140. The method of claim 139 wherein the diseased
25 condition is a malignancy.

141. The method of claim 123 wherein said DNA sample is cDNA prepared from mRNA.

30   142. A method for identifying, classifying, or quantifying DNA molecules in a sample of DNA molecules with a plurality of nucleotide sequences, the method comprising the steps of:
(a) digesting said sample with one or more restriction
35 endonucl as s, each said restriction endonuclease recognizing a subsequenc recognition site and digesting DNA to produce fragm nts with 3' overhangs;

- 286 -

(b)   contacting said fragments with shorter and long r oligodeoxynucleotides, each said longer oligodeoxynucleotide consisting of a first and second contiguous portion, said first portion being a 3' end subsequence complementary to the
5 overhang produced by one of said restriction endonucleases, each said shorter oligodeoxynucleotide complementary to the 3' end of said second portion of said longer oligodeoxynucleotide stand;

(c)   ligating said longer oligodeoxynucleotides to said
10 DNA fragments to produce a ligated fragments;

(d)   extending said ligated DNA fragments by synthesis with a DNA polymerase to form blunt-ended double stranded DNA fragments;

(e)   amplifying said double stranded DNA fragments by
15 use of a DNA polymerase and primer oligodeoxynucleotides to produce amplified DNA fragments, each said primer oligodeoxynucleotide having a sequence comprising that of a longer oligodeoxynucleotide;

(f)   determining the length of the amplified DNA
20 fragments; and

(g)   searching a DNA sequence database, said database comprising a plurality of known DNA sequences that may be present in the sample, for sequences matching one or more of said fragments of determined length, a sequence from said
25 database matching a fragment of determined length when the sequence from said database comprises recognition sites of said one or more restriction endonucleases spaced apart by the determined length,

whereby DNA sequences in said sample are identified,
30 classified, or quantified.

143.   A method of detecting one or more differentially expressed genes in an *in vitro* cell exposed to an exogenous factor relative to an *in vitro* cell not exposed to said
35 exog nous factor comprising:

(a)   performing the m thod of claim 1 wherein said
plurality of nucleic acids comprises cDNA of RNA of said *in
vitro* cell exposed to said exogenous factor;

(b)   performing the method of claim 1 wherein said
5 plurality of nucleic acids comprises cDNA of RNA of said *in
vitro* cell not exposed to said exogenous factor; and

(c)   comparing the identified, classified, or quantified
cDNA of said *in vitro* cell exposed to said exogenous factor
with the identified, classified, or quantified cDNA of said
10 *in vitro* cell not exposed to said exogenous factor,

whereby differentially expressed genes are identified,
classified, or quantified.

144.   A method of detecting one or more differentially
15 expressed genes in a diseased tissue relative to a tissue not
having said disease comprising:

(a)   performing the method of claim 1 wherein said
plurality of nucleic acids comprises cDNA of RNA of said
diseased tissue, such that one or more cDNA molecules are
20 identified, classified, and/or quantified;

(b)   performing the method of claim 1 wherein said
plurality of nucleic acids comprises cDNA of RNA of said
tissue not having said disease, such that one or more cDNA
molecules are identified, classified, and/or quantified; and
25   (c)   comparing said identified, classified, and/or
quantified cDNA molecules of said diseased tissue with said
identified, classified, and/or quantified cDNA molecules of
said tissue not having the disease,

whereby differentially expressed cDNA molecules are
30 detected.

145.   The method of claim 144 wherein the step of
comparing further comprises finding cDNA molecules which are
reproducibly expressed in said diseased tissue or in said
35 tissue not having the disease and further finding which of
said reproducibly expr ssed cDNA molecules have significant

differences in expression between the tissue having said
disease and the tissue not having said disease.

146. The method of claim 145 wherein said finding cDNA
5 molecules which are reproducibly expressed and said
significant differences in expression of said cDNA molecules
in said diseased tissue and in said tissue not having the
disease are determined by a method comprising applying
statistical measures.
10

147. The method of claim 146 wherein said statistical
measures comprise finding reproducible expression if the
standard deviation of the level of quantified expression of a
cDNA molecule in said diseased tissue or said tissue not
15 having the disease is less than the average level of
quantified expression of said cDNA molecule in said diseased
tissue or said tissue not having the disease, respectively,
and wherein a cDNA molecule has significant differences in
expression if the sum of the standard deviation of the level
20 of quantified expression of said cDNA molecule in said
diseased tissue plus the standard deviation of the level of
quantified expression of said cDNA molecule in said tissue
not having the disease is less than the absolute value of the
difference of the level of quantified expression of said cDNA
25 molecule in said diseased tissue minus the level of
quantified expression of said cDNA molecule in said tissue
not having the disease.

148. The method of claim 144 wherein the diseased
30 tissue and the tissue not having the disease are from one or
more mammals.

149. The method of claim 144 wherein the disease is a
malignancy.
35

150. The method of claim 144 wherein the disease is a
malignancy select d from the group consisting of prostrate

canc r, br ast cancer, colon canc r, lung cancer, skin
cancer, lymphoma, and leukemia.

151.   The method of claim 144 wherein the disease is a
5 malignancy and the tissue not having the disease has a
premalignant character.

152.   A method of staging or grading a disease in a
human individual comprising:
10   (a) performing the method of claim 1 in which said
plurality of nucleic acids comprises cDNA of RNA prepared
from a tissue from said human individual, said tissue having
or suspected of having said disease, whereby one or more said
cDNA molecules are identified, classified, and/or quantified;
15 and
       (b) comparing said one or more identified, classified,
and/or quantified cDNA molecules in said tissue to the one or
more identified, classified, and/or quantified cDNA molecules
expected at a particular stage or grade of said disease.
20
       153.   A method for predicting a human patient's response
to therapy for a disease, comprising:
       (a) performing the method of claim 1 in which said
plurality of nucleic acids comprises cDNA of RNA prepared
25 from a tissue from said human patient, said tissue having or
suspected of having said disease, whereby one or more cDNA
molecules in said sample are identified, classified, and/or
quantified; and
       (b) ascertaining if the one or more cDNA molecules
30 thereby identified, classified, and/or quantified correlates
with a poor or a favorable response to one or more therapies.

       154.   The method of claim 153 which further comprises
selecting one or more therapies for said patient for which
35 said identified, classified, and/or quantified cDNA molecules
corr lates with a favorable r sponse.

155. A method for evaluating the efficacy of a therapy
in a mammal having a dis ase, the method comprising:

(a)   performing the method of claim 1 wherein said
plurality of nucleic acids comprises cDNA of RNA of said
5 mammal prior to a therapy;

(b)   performing the method of claim 1 wherein said
plurality of nucleic acids comprises cDNA of RNA of said
mammal subsequent to said therapy;

(c)   comparing one or more identified, classified,
10 and/or quantified cDNA molecules of said mammal prior to said
therapy with one or more identified, classified, and/or
quantified cDNA molecules of said mammal subsequent to
therapy; and

(d)   determining whether the response to therapy is
15 favorable or unfavorable according to whether any differences
in the one or more identified, classified, and/or quantified
cDNA molecules after therapy are correlated with regression
or progression, respectively, of the disease.

20   156. The method of claim 155 wherein the mammal is a
human.

157. A kit comprising:

(a)   one or more containers having one or more
25 restriction endonucleases;

(b)   one or more containers having one or more shorter
oligodeoxynucleotide strands;

(c)   one or more containers having one or more longer
oligodeoxynucleotide strands hybridizable with said shorter
30 strands, wherein either the longer or the shorter
oligodeoxynucleotide strands each comprise a subsequence
complementary to a single-stranded overhang produced by at
least one of said one or more restriction endonucleases; and

(d)   instructions packaged in association with said one
35 or mor containers for use of said restriction endonucl as s,
short r strands, and longer strands for identifying,

classifying, or quantifying one or more DNA molecules in a
DNA sample, said instructions comprising:

      i. digest said sample with said restriction
endonucleases into fragments, each fragment being terminated
5 on each end by a single-stranded overhang of said one or more
restriction endonucleases;

      ii. contact said shorter and longer strands and
said digested fragments to form double stranded DNA adapters
annealed to said digested fragments,

10   iii. ligate said longer strand to said fragments;

      iv. generate one or more signals by separating and
detecting such of said fragments that are digested on each
end, each signal comprising a representation of the length of
the fragment and the identity of the recognition sites on
·15 both termini of the fragments; and

      v. search a nucleotide sequence database to
determine sequences that match or the absence of any
sequences that match said one or more generated signals, said
database comprising a plurality of known nucleotide sequences
20 of nucleic acids that may be present in the sample, a
sequence from said database matching a generated signal when
the sequence from said database has both (i) the same length
between occurrences of said recognition sites of said one or
more restriction endonucleases as is represented by the
25 generated signal and (ii) the same recognition sites of said
one of more restriction endonucleases as is represented by
the generated signal.

    158. The kit of claim 157 wherein said one or more
30 restriction endonucleases generate 5' overhangs at the
terminus of digested fragments, wherein each said shorter
oligodeoxynucleotide strand consists of a first and second
contiguous portion, said first portion being a 5' end
subsequence complementary to the overhang produced by one of
35 said restriction endonucleases, and wher in ach said longer
oligod oxynucleotide strand comprises a 3' end subsequ nce

complementary to said second portion of said shorter
oligod oxynucleotide strand.

159. The kit of claim 157 wherein said one or more
5 restriction endonucleases generate 3' overhangs at the
terminus of the digested fragments, wherein each said longer
oligodeoxynucleotide strand consists of a first and second
contiguous portion, said first portion being a 3' end
subsequence complementary to the overhang produced by one of
10 said restriction endonucleases, and wherein each said shorter
oligodeoxynucleotide strand is complementary to the 3' end of
said second portion of said longer oligodeoxynucleotide
stand.

15   160. The kit of claim 157 wherein said instructions
further comprise those signals expected from one or more DNA
molecules of interest when said sample is digested with a
particular one or more restriction endonucleases selected
from among said one or more restriction endonucleases in said
20 kit.

161. The kit of claim 160 wherein said one or more DNA
molecules of interest are cDNA molecules differentially
expressed in a disease condition.
25

162. The kit of claim 157 wherein the restriction
endonucleases are selected from the group consisting of
Acc65I, AflII, AgeI, ApaLI, ApoI, AscI, AvrI, BamHI, BclI,
BglII, BsiWI, Bsp120I, BspEI, BspHI, BsrGI, BssHII, BstYI,
30 EagI, EcoRI, HindIII, MluI, NcoI, NgoMI, NheI, NotI, SpeI,
and XbaI.

163. The kit of claim 157 which comprises one or more
containers having one or more double stranded adapter DNA
35 molecules formed by annealing said longer and said shorter
oligonucleotide strands.

164.   The kit of claim 157 further comprising a computer readable memory according to claim 106.

165.   The kit of claim 157 further comprising a computer
5 readable memory according to claim 114.

· 166.   The kit of claim 157 further comprising a computer readable memory according to claim 122.

10   167.   The kit of claim 157 further comprising in a container a DNA ligase.

168.   The kit of claim 157 further comprising in a container a phosphatase capable of removing terminal
15 phosphates from a DNA sequence.

169.   The kit of claim 157 further comprising in one or more containers:
        (a)   one or more primers, each said primer consisting of
20 a single stranded oligodeoxynucleotide comprising the sequence of one of said longer strands; and
        (b)   a DNA polymerase.

170.   The kit of claim 169 wherein each of said one or
25 more primers further comprises (a) a first subsequence that is the portion of the recognition site of one of said one or more restriction endonucleases remaining at the terminus of a fragment after digestion, and (b) a second subsequence of one or two additional nucleotides contiguous with and 3' to said
30 first subsequence, wherein said primer is detectably labeled such that primers with differing said one or two additional nucleotides have different labels that can be distinguishably detected.

35   171.   The kit of claim 157 wh r in said instructions furth r comprise: det ct such of said fragments digested on each end by a method comprising staining said fragments with

- 294 -

silver, labeling said fragments with a DNA int rcalating dye,
or det cting light emission from a fluorochrome label on said
fragments.

5   172. The kit of claim 157 further comprising:
        (a) reagents for performing a cDNA sample preparation
step;
        (b) reagents for performing a step of digestion by one
or more restriction endonucleases;
10   (c) reagents for performing a ligation step; and
        (d) reagents for performing a PCR amplification step.

        173. A method for identifying, classifying, or
quantifying one or more nucleic acids in a sample comprising
15 a plurality of nucleic acids having different nucleotide
sequences, said method comprising:
        (a) probing said sample with one or more recognition
means, each recognition means causing recognition of a target
nucleotide subsequence or a set of target nucleotide
20 subsequences;
        (b) generating one or more signals from said sample
probed by said recognition means, each generated signal
arising from a nucleic acid in said sample and comprising a
representation of (i) the identities of effective
25 subsequences, each said effective subsequence being a
subsequence comprising a target subsequence, or the
identities of sets of effective subsequences, each said set
having member effective subsequences each of which comprises
a different target subsequence from one of said sets of
30 target sequences, and (ii) the length between occurrences of
effective subsequences in said nucleic acid or between one
occurrence of one effective subsequence and the end of said
nucleic acid; and
        (c) searching a nucleotide sequence database to
35 determine sequu nces that match or the absence of any
sequenc s that match said one or more generated signals, said
database comprising a plurality of known nucleotid sequences

- 295 -

of nucleic acids that may be present in the sample, a
sequence from said database matching a generated signal when
the sequence from said database has both (i) the same length
between occurrences of effective subsequences or the same
5 length between one occurrence of one effective target
subsequence and the end of the sequence as is represented by
the generated signal, and (ii) the same effective
subsequences as are represented by the generated signal, or
effective subsequences that are members of the same sets of
10 effective subsequences as are represented by the generated
signal,

   whereby said one or more nucleic acids in said sample
are identified, classified, or quantified.

15   174.   The method according to claim 173 wherein said one
or more nucleic acids are DNA.

       175.   The method according to claim 174 wherein one or
more of said effective subsequences consist of a single
20 target nucleotide subsequence.

       176. The method according to claim 175 wherein each of
said single target nucleotide subsequences is recognized by a
first restriction endonuclease having said single target
25 nucleotide subsequence for its recognition site.

       177.   The method according to claim 174 wherein one or
more of said effective subsequences consist of two target
nucleotide subsequences, a first target nucleotide
30 subsequence and a second target nucleotide subsequence.

       178. The method according to claim 177 wherein each of
said first target nucleotide subsequences is recognized by a
restriction endonuclease having said first target nucleotide
35 subsequ nce for its recognition site.

179. The method according to claim 177 wher in each of said second target nucleotide subsequences is at least a portion of a single-stranded overhang produced by a Type IIS restriction endonuclease.

5

180. The method according to claim 177 wherein said first target nucleotide subsequence and said second target nucleotide subsequence are adjacent in said one or more said effective subsequences.

10

181. The method according to claim 174 further comprising:

(a) identifying a fragment of a nucleic acid in said sample which generates one or more of said signals; and

15 (b) recovering said fragment.

182. The method according to claim 181 which further comprises using at least a hybridizable portion of said fragment as a hybridization probe to hybridize to a nucleic

20 acid that comprises said fragment.

183. The method according to claim 174 wherein one or more of said signals do not match any sequence in said nucleotide sequence database.

25

184. The method according to claim 174 wherein said DNA is cDNA synthesized from source mRNA according to a synthesis method comprising using one or more primers with a conjugated capture moiety.

30

185. The method according to claim 174 wherein said DNA is cDNA synthesized from source mRNA, and wherein said end of said nucleic acid has a fixed offset from a 5'-cap of said source mRNA.

35

186. The method according to claim 185 wherein said synthesis is according to a method comprising a step of

ligating a DNA-RNA chimera to said source mRNA in a fixed
offs t to said 5'-cap of said source mRNA.

187.   The method according to claim 186 wherein said
5 ligating in a fixed offset comprises ligating to the
ribonucleotide 3' adjacent to the 5' cap ribonucleotide.

188.   The method according to claim 186 wherein said
DNA-RNA chimera comprises a DNA portion having a conjugated
10 capture moiety.

189.   The method according to claim 188 wherein said
probing step further comprises a step of digesting said
sample into fragments and subsequent steps of contacting said
15 fragments with a binding partner of said capture moiety
affixed to a solid support, washing said binding partner, and
denaturing DNA bound to said binding partner to release
strands not having a conjugated capture moiety, and wherein
said generating step generates signals from said released
20 strands.

190. The method according to claim 174 wherein said DNA
is cDNA synthesized from source mRNA, and wherein said end of
said nucleic acid has a fixed offset from the 5' end of the
25 3' poly(A) tail of said source mRNA.

191.   The method according to claim 190 wherein said
synthesis is according to a method comprising a step of
synthesizing a cDNA strand with a nucleotide polymerase
30 primed by oligo(dT) phasing primers.

192.   The method according to claim 174 wherein one or
more of said recognition means are Type IIS restriction
endonucleases, wherein the step of probing further comprises
35 a first digesting step in which said sample is dig sted with
said one or more Type IIS restriction ndonucl ases, th reby
cr ating first singl -stranded ov rhangs on cut fragments

outside of the recognition sites of the Type IIS restriction
endonucleases, and wherein one or more of said effective
subsequences comprise a first target nucleotide subsequence
that is the sequence of at least a portion of one of said
5 first single-stranded overhangs.

193.   The method according to claim 174 wherein one or
more of said recognition means are restriction endonucleases,
wherein the step of probing further comprises a first
10 digesting step in which said sample is digested with said one
or more restriction endonucleases, thereby creating first
single-stranded overhangs on cut fragments, and wherein one
or more of said effective subsequences comprise a first
target nucleotide subsequence that is a recognition site of
15 one of said restriction endonucleases.

194.   The method according to claim 193 wherein said
step of probing further comprises after said digesting step:
        a step of hybridizing partially double-stranded adapter
20 nucleic acids with said fragments, each said adapter nucleic
acid comprising a primer strand and a linker strand, each
said linker oligonucleotide (i) being shorter in sequence
than said primer oligonucleotide, and (ii) having an end
complementary to one of said first single-stranded overhangs;
25 and
        a step of ligating said primer strand of said hybridized
adapter nucleic acid to said first single-stranded overhang,
whereby ligated fragments are formed.

30   195.   The method according to claim 194 wherein said
step of probing further comprises after said steps of
hybridizing and ligating a step of amplifying said ligated
fragments, whereby amplified fragments are formed.

35   196.   The method according to claim 195 wh rein said
amplifying is carri d out by use of a nucleic acid polymerase
and primers that are singl -stranded nucleic acids comprising

the sequenc of said primer strands of said adapter nucleic acids, said primers being capable of priming nucleic acid synthesis by said polymerase.

5   197.   The method according to claim 196 wherein said primers further comprise a conjugated capture moiety and a release means, and wherein said step of probing further comprises after said step of amplifying a step of cleanup, which comprises contacting said amplified fragments with a

10 binding partner of said capture moiety affixed to a solid support, washing said binding partner in denaturing conditions, and using said release means to release fragments bound to said binding partner, whereby released fragments are formed, and wherein said generating step generates signals

15 from said released fragments.

       198.   The method according to claim 196 wherein said primers further comprise a conjugated capture moiety, and wherein said step of probing further comprises after said

20 step of amplifying a step of cleanup, which comprises contacting said amplified fragments with a binding partner of said capture moiety affixed to a solid support, washing said binding partner, and denaturing DNA bound to said binding partner to release fragments not having a conjugated capture

25 moiety, whereby released fragments are formed, and wherein said generating step generates signals from said released fragments.

       199. The method according to claim 198 wherein said

30 capture moiety is biotin.

       200. The method according to claim 198 wherein said step of generating further comprises separating said released fragments according to length and detecting the lengths of

35 said fragments.

201. The method according to claim 200 wherein said separating is done by electrophoresis in denaturing conditions.

5   202. The method according to claim 194 wherein said step of generating further comprises a step of separating said ligated fragments according to length and detecting the lengths of such fragments.

10   203. The method according to claim 202 wherein said primer strands of said adapter nucleic acids are distinguishably labeled, and wherein said step of generating further comprises detecting said labels of said primer strands.

15

204. The method according to claim 202 wherein prior to said generating step said ligated fragments are amplified.

205. The method according to claim 194 wherein one or
20 more of said recognition means are Type IIS restriction endonucleases, wherein the step of probing further comprises a second digesting step in which said sample is digested with said one or more Type IIS restriction endonucleases, thereby creating second single-stranded overhangs on cut fragments
25 outside of the recognition sites of the Type IIS restriction endonucleases, and wherein one or more of said effective subsequences comprise a second target nucleotide subsequence that is the sequence of at least a portion of one of said second single-stranded overhangs.

30

206. The method according to claim 205 wherein said primer strands of said adapter nucleic acids comprise a recognition site for a Type IIS restriction endonuclease.

35   207. The m thod according to claim 205 wherein said primer strands of said adapter nucleic acids comprise a recognition site for a Type IIS restriction endonuclease so

positioned that said first target nucleotide subsequence and
said second target nucleotide subsequ nce are adjacent in
said one or more said effective subsequences.

5   208.   The method according to claim 205 wherein the step
of probing further comprises before said second digesting
step a step of amplifying said ligated fragments.

       209.   The method according to claim 205 wherein said
10 primer strands further comprise a conjugated capture moiety,
and wherein the step of probing further comprises before said
second digestion step a step which comprises contacting said
ligated fragments with a binding partner of said capture
moiety affixed to a solid support, washing said binding
15 partner, and denaturing DNA bound to said binding partner to
release fragments not having a conjugated capture moiety,
whereby released fragments are formed, and wherein said
generating step generates signals from said released
fragments.
20

       210.   The method according to claim 205 wherein said
generating step further comprises a step of determining the
sequences of said second single-stranded overhangs.

25   211.   The method according to claim 210 wherein said
step of determining the sequence comprises performing Sanger
sequencing reactions on the second single-stranded overhangs
primed by the shorter strand of said fragments with said
second single-stranded overhangs.
30

       212.   The method according to claim 211 wherein said
primer strand ligated to said shorter strand of said
fragments with said second single-stranded overhangs
comprises a release means, and wherein said step of
35 det rmining th s quence further compris s after said
performing Sanger sequencing r actions release of said
short r strand by said r lease means.

213. Th method according to claim 212 wherein said release means comprises a subsequence of one or more uracil nucleotides or a recognition site for a rare cutting restriction endonuclease.

214. The method according to claim 213 wherein said rare cutting restriction endonuclease is AscI.

215. The method according to claim 210 wherein said step of determining the sequence comprises performing a method comprising:

(a) hybridizing partially double-stranded second adapter nucleic acids to said second single-stranded overhangs, each said second adapter nucleic acid comprising a second primer oligonucleotide strand and a second phasing linker oligonucleotide strand, each said second phasing linker oligonucleotide strand being shorter in sequence than said second primer oligonucleotide strand, and having one fixed nucleotide hybridizable with one position of said second single-stranded overhangs and random nucleotides hybridizable with remaining positions of said second single-stranded overhangs, said second primer oligonucleotide strand being distinctively labeled according to the identity of said one fixed nucleotide;

(b) ligating said second primer oligonucleotide strands to said overhangs with a ligase;

(c) detecting which said second primer oligonucleotide strand has been ligated to said overhangs and thereby determining which fixed nucleotide hybridizes with said one position of said second single-stranded overhang; and

(d) repeating steps (a) through (c) for all possible nucleotides at all possible positions of said second single-stranded overhang;

wh r by the nucl otide sequenc of said second single-stranded overhang is determined.

216.   A m thod for identifying, classifying, or
quantifying DNA molecules in a sample comprising DNA
molecules having a plurality of different nucleotide
sequences, the method comprising the steps of:

5   (a)   digesting said sample with one or more first
restriction endonucleases, each said first restriction
endonuclease cutting DNA molecules in said sample at its
recognition site to produce fragments with first single-
stranded overhangs of known sequence;

10   (b)   hybridizing said fragments with linker
oligonucleotides and first primer oligonucleotides, each said
linker oligonucleotide hybridizable with one of said first
single-stranded overhangs, and each said first primer
oligonucleotide hybridizable with one of said linker
15 oligonucleotides;

        (c) ligating said first primer oligonucleotides to the
end of said first single-stranded overhangs, whereby ligated
fragments are formed;

        (d)   amplifying said ligated fragments to produce
20 amplified fragments by a method comprising contacting said
ligated fragments with a DNA polymerase and second primer
oligonucleotides, each said second primer oligonucleotide
having a sequence comprising that of one of said first primer
oligonucleotides, and wherein one of said second primer
25 oligonucleotides has a conjugated capture moiety, and one of
said second primer oligonucleotides comprises a recognition
site for a Type IIS restriction endonuclease, whereby
amplified fragments are formed;

        (e)   binding said amplified fragments by contacting said
30 amplified fragments with the binding partner of said capture
moiety affixed to a solid support, whereby bound fragments
are formed;

        (f)   washing at least a portion of said bound fragments;

        (g)   denaturing at least a portion of said bound washed
35 fragments to release strands not conjugated to said capture
moiety;

- 304 -

(h)   determining the length of at least one of said
released strands, thereby determining one or more fragments
of determined length;

(i)   digesting at least a portion of said bound, washed
5 fragments with said Type IIS restriction endonuclease to
produce fragments with second single-stranded overhangs; and

(j)   sequencing said second single-stranded overhangs
created in step (i).

10   217.   The method according to claim 216 wherein said
recognition site for said Type IIS restriction endonuclease
is so positioned that upon cleavage with said Type IIS
restriction endonuclease one of said second single-stranded
overhangs is created adjacent to the recognition site of said
15 first restriction endonuclease, and said method further
comprising, after said sequencing step, a step of searching a
DNA sequence database for sequences matching or the absence
of any sequences matching one or more of said fragments of
determined length, said database comprising a plurality of
20 known DNA sequences that may be present in the sample, a
sequence from said database matching a fragment of determined
length when the sequence from said database comprises
effective subsequences spaced apart by said determined
length, said effective subsequences consisting of said
25 recognition sites of said one or more first restriction
endonucleases concatenated to sequences of said second
single-stranded overhangs.

218.   The method according to claim 216 wherein said
30 recognition site for said Type IIS restriction endonuclease
is so positioned that upon cleavage with said Type IIS
restriction endonuclease one of said second single-stranded
overhangs is created non-contiguous with the recognition site
of said first restriction endonuclease, and said method
35 further comprising, after said sequ ncing step, a step of
searching a DNA sequence database for sequences matching or
the absence of any sequences matching on or more of said

fragments of determined length, said database comprising a
plurality of known DNA s quences that may be present in the
sample, a sequence from said database matching a fragment of
determined length when the sequence from said database
5 comprises the same recognition sites of said one or more
first restriction endonucleases spaced apart by said
determined length.

    219. The method according to claim 218 wherein a
10 sequence from said database matches a fragment of determined
    length only when said sequence from said database further
    comprises the sequence of said second single-stranded
    overhang spaced apart from said recognition site in said
    sequence by the same number of nucleotides as the second
15 single-stranded overhang is spaced apart from said
    recognition site in said fragment of determined length.

    220.   A method for identifying, classifying, or
quantifying cDNA molecules in a sample comprising cDNA
20 molecules having a plurality of different nucleotide
    sequences, the method comprising the steps of:
        (a)   digesting said sample with one or more first
    restriction endonucleases, each said first restriction
    endonuclease cutting cDNA molecules in said sample at its
25 recognition site to produce fragments with single-stranded
    overhangs of known sequence;
        (b)   hybridizing said fragments with linker
    oligonucleotides and first primer oligonucleotides, each said
    linker oligonucleotide hybridizable with one of said single-
30 stranded overhangs, and each said first primer
    oligonucleotide hybridizable with one of said linker
    oligonucleotides;
        (c) ligating said first primer oligonucleotides to the
    end of said single-stranded overhangs, whereby ligated
35 fragm nts are formed;
        (d)   amplifying said ligated fragments to produce
    amplifi d fragments by a m thod comprising contacting said

- 306 -

ligated fragments with a DNA polym rase and second primer
oligonucl otid s, one of said second primer oligonucleotides
having a conjugated capture moiety and a sequence comprising
that of a 5'-cap-primer oligonucleotide, and each of the
5 remaining of said second primer oligonucleotides having a
sequence comprising that of one of said first primer
oligonucleotides, wherein said sample of cDNA was synthesized
from source mRNA by a method such that cDNA molecules in said
sample comprise said 5'-cap-primer oligonucleotide in a fixed
10 relation to the 5'-cap of said source mRNA; whereby amplified
fragments are formed;

     (e)   binding said amplified fragments by contacting said
amplified fragments with the binding partner of said capture
moiety affixed to a solid support, whereby bound fragments
15 are formed;

     (f)   washing at least a portion of said bound fragments;

     (g)   denaturing at least a portion of said bound, washed
fragments to release strands not conjugated to said capture
moiety; and

20   (h)   determining the length of at least one of said
released strands, thereby determining one or more fragments
of determined length.

     221.   A method for identifying, classifying, or
25 quantifying DNA molecules in a sample comprising DNA
molecules having a plurality of different nucleotide
sequences, the method comprising the steps of:

     (a)   digesting said sample with one or more first
restriction endonucleases, each said first restriction
30 endonuclease cutting DNA molecules in said sample at its
recognition site to produce fragments with single-stranded
overhangs of known sequence;

     (b)   hybridizing said fragments with linker
oligonucleotides and first primer oligonucleotides, each said
35 linker oligonucl otid hybridizable with one of said singl -
stranded overhangs, and each said first primer

oligonucl otid hybridizable with one of said linker
oligonucleotides;

      (c) ligating said first primer oligonucleotides to the
end of said single-stranded overhangs; whereby ligated
5 fragments are formed;

      (d)   amplifying said ligated fragments to produce
amplified fragments by a method comprising contacting said
ligated fragments with a DNA polymerase and second primer
oligonucleotides, each said second primer oligonucleotide
10 having a sequence comprising that of one of said first primer
oligonucleotides, and wherein one of said second primer
oligonucleotides comprises a conjugated capture moiety;
whereby amplified fragments are formed;

      (e)   binding said amplified fragments by contacting said
15 amplified fragments with the binding partner of said capture
moiety affixed to a solid support; whereby bound fragments
are formed;

      (f)   washing at least a portion of said bound fragments;

      (g)   denaturing at least a portion of said bound, washed
20 fragments to release strands not conjugated to said capture
moiety; and

      (h)   determining the length of at least one of said
released strands, thereby determining one or more fragments
of determined length.
25

      222.   The method according to claim 221 wherein said
linker oligonucleotide is shorter in sequence that said first
primer oligonucleotide.

30   223.   The method according to claim 221 further
comprising prior to said hybridizing step a step of
hybridizing said linker oligonucleotides and said first
primer oligonucleotides to form partially double stranded
adapter nucleic acids, and wherein said hybridizing step
35 comprises hybridizing said adapter nucleic acids to said
ov rhangs.

224. The method according to claim 221 further
comprising prior to said amplifying step a step of extending
said ligated DNA fragments to blunt-ended double stranded DNA
fragments by synthesis with a DNA polymerase.

5

225. The method according to claim 221 wherein the
steps of digesting, hybridizing, and ligating are performed
simultaneously in a first reaction mix, and wherein the step
of amplifying is performed in a second reaction mix.

10

226. The method according to claim 225 wherein said
first and said second reaction mixes are in the same reaction
vessel separated by a separation means during said steps of
digesting, hybridizing, ligating, and amplifying, and wherein

15 the step of amplifying further comprises a step of causing
said first and said second reaction mixes to combine across
said separation means.

227. The method according to claim 226 wherein said
20 separation means is a wax, and wherein said amplifying step
melts said wax.

228. The method according to claim 227 wherein said wax
is a mixture of Paraffin and Chillout™ waxes.

25

229. The method according to claim 221 wherein said
determining step further comprises separating said released
strands according to length by electrophoresis.

30   230. The method according to claim 221 wherein said
determining step further comprises detecting said release
strands by a method comprising labeling said fragments with a
DNA intercalating dye or Ag staining or detecting light
emission from a fluorochrome label on said fragments.

35

231. The m thod according to claim 221 wherein said ne
or more first r striction endonucleases are pairs of

restriction endonucleases, the pairs being selected from the
group consisting of Acc56I and HindIII, Acc65I and NgoMI,
BamHI and EcoRI, BglII and HindIII, BglII and NgoMI, BsiWI
and BspHI, BspHI and BstYI, BspHI and NgoMI, BsrGI and EcoRI,   -
5 EagI and EcoRI, EagI and HindIII, EagI and NcoI, HindIII and
NgoMI, NgoMI and NheI, NgoMI and SpeI, BglII and BspHI,
Bsp120I and NcoI, BssHII and NgoMI, EcoRI and HindIII, and
NgoMI and XbaI.

10   232.   The method according to claim 221 wherein the step
of ligating is performed with T4 DNA ligase.

         233.   The method according to claim 221 wherein said
capture moiety is biotin.
15

         234.   The method according to claim 221 further
comprising after said determining step a step of searching a
DNA sequence database for sequences matching or the absence
of any sequences matching one or more of said fragments of
20 determined length, said database comprising a plurality of
known DNA sequences that may be present in the sample, a
sequence from said database matching a fragment of determined
length when the sequence from said database comprises the
same recognition sites of said one or more restriction
25 endonucleases present in the fragment and spaced apart by the
same determined length.

         235.   The method according to claim 234 wherein the step
of searching said DNA database further comprises:
30   (a)   determining a pattern of simulated fragments that
can be generated from sequences in said DNA database and for
each simulated fragment in said pattern those sequences in
said DNA database that are capable of generating the fragment
by simulating the steps of digesting with said one or more
35 restriction endonucleas s, hybridizing, ligating, amplifying,
and d t rmining appli d to each sequenc in said DNA
database; and

(b) finding the sequences in said database that are
capable of generating said one or more fragments of
determined length by finding in said pattern one or more
fragments that have the same recognition sites spaced apart
5 by the same length as said one or more fragments of
determined length.

236. The method according to claim 221 wherein the
steps of digesting and ligating go substantially to
10 completion.

237. The method according to claim 221 wherein the DNA
sample is cDNA of RNA from a tissue or a cell type derived
from a plant, a single celled animal, a multicellular animal,
15 a bacterium, a virus, a fungus, or a yeast.

238. A method of detecting the presence of
differentially expressed cDNAs in a first tissue relative to
a second tissue comprising:
20   (a) performing the method of claim 218 wherein said
sample of DNA molecules comprises cDNA of RNA of said first
tissue, and wherein lengths are determined for one or more
released strands from said first tissue;
    (b) performing the method of claim 218 wherein said
25 sample of DNA molecules comprises cDNA of RNA of said second
tissue, and wherein lengths are determined for one or more
released strands from said second tissue; and
    (c) comparing said released strands of determined
length from said first tissue with said released strands of
30 determined length from said second tissue.
    whereby the presence of differentially expressed cDNA
molecules are detected.

239. The method according to claim 238 wherein the step
35 f comparing further comprises finding which of said releas d
strands which ar r pr ducibly expressed in said first tissue
or in said second tissu and further finding which of said

reproducibly express d released strands have significant differences in expression between said first tissue and said second tissue.

5   240. The method according to claim 239 further comprising:
        (a) identifying said released strands having significant differences in expression; and
        (b) recovering said differently expressed fragment.
10

        241. The method according to claim 240 which further comprises using at least a hybridizable portion of said released strands having significant differences in expression as a hybridization probe to bind to a nucleic acid that can
15 generate said released strands having significant differences in expression according to the method of claim 218.

        242. The method according to claim 239 wherein said finding released strands which are reproducibly expressed and
20 said finding significant differences in expression of said released strands in said first tissue and in said second tissue are determined by a method comprising applying statistical measures.

25   243. The method according to claim 242 wherein said statistical measures comprise finding reproducible expression if the standard deviation of the level of quantified expression of a released strand in said first tissue or said second tissue is less than the average level of quantified
30 expression of said released strand in said first tissue or said second tissue, respectively, and wherein a released strand has significant differences in expression if the sum of the standard deviation of the level of quantified expression of said released strand in said first tissue plus
35 th standard d viations of the level of quantified expression of said r l ased strand in said s cond tissue is less than the absolute valu of the diff r nce f the level of

quantified xpression of said released strand in said first
tissue minus the level of quantified expression of said
released strand in said second tissue.

5   244. The method according to claim 238 wherein said
first tissue and said second tissue are different tissue-
types from the same organism.

245. The method according to claim 238 wherein said
10 first tissue and said second tissue are the same tissue-type
from phylogenetically related organisms.

246. The method according to claim 238 wherein said
first tissue and said second tissue are the same tissue-type
15 from an organism in a first condition and said organism in a
second condition.

247. The method according to claim 246 wherein said
first condition is a normal condition and said second
20 condition is a diseased condition.

248. A kit comprising:
         (a) one or more containers having one or more
restriction endonucleases;
25   (b) one or more containers having one or more shorter
oligodeoxynucleotide strands;
         (c) one or more containers having one or more longer
oligodeoxynucleotide strands hybridizable with said shorter
strands, wherein either the longer or the shorter
30 oligodeoxynucleotide strands each comprise a sequence
complementary to an overhang produced by at least one of said
one or more restriction endonucleases, and wherein one or
more of said longer oligodeoxynucleotide strands have a
conjugated capture moiety; and
35   (d) instructions packaged in association with said one
or more containers for us of said r striction endonucleases,
shorter strands, and longer strands for identifying,

classifying, or quantifying one or more DNA mol cules in a
DNA sample, said instructions comprising:

       i.   digest said sample with said restriction
endonucleases into fragments, each fragment being terminated
5 on each end by a recognition site of said one or more
restriction endonucleases;

       ii.   contact said shorter and longer strands and
said digested fragments to form double stranded DNA adapters
annealed to said digested fragments,

10   iii.   ligate said longer strand to said digested
fragments such that each digested fragment has ligated to it
one of said longer strands with a conjugated capture moiety;

       iv.   contact said digested fragments to a binding
partner of said capture moiety, wash, and denature single
15 strands not conjugated to a capture moiety from said binding
partner;

       v.   generate one or more signals from said
denatured single strands, each signal comprising a
representation of the length of the fragment and the identity
20 of the recognition sites on both termini of the fragments;

       vi.   search a nucleotide sequence database to
determine sequences that match or the absence of any
sequences that match said one or more generated signals, said
database comprising a plurality of known nucleotide sequences
25 of nucleic acids that may be present in the sample, a
sequence from said database matching a generated signal when
the sequence from said database has both (1) the same length
between occurrences of said recognition sites of said one or
more restriction endonucleases as is represented by the
30 generated signal, and (2) the same recognition sites of said
one of more restriction endonucleases as is represented by
the generated signal.

     249.   The kit of claim 248 which comprises one or more
35 c ntainers having one or more double stranded adapter DNA
molecules form d by ann aling said long r and said shorter
oligonucl otid strands.

250.   A kit comprising:

(a)   one or mor containers having one or more non-Type IIS restriction endonucleases;

(b)   one of more containers having one or more Type IIS
5 restriction endonucleases;

(c)   one or more containers having one or more shorter oligodeoxynucleotide strands;

(d)   one or more containers having one or more longer oligodeoxynucleotide strands hybridizable with said shorter
10 strands, wherein either the longer or the shorter oligodeoxynucleotide strands each comprise a sequence complementary to an overhang produced by at least one of said one or more non-Type IIS restriction endonucleases, and wherein one or more of said longer oligodeoxynucleotide
15 strands has a recognition site for one of said Type IIS restriction endonucleases; and

(e)   instructions packaged in association with said one or more containers for use of said non-Type IIS restriction endonucleases, Type IIS restriction endonucleases, shorter
20 strands, and longer strands for identifying, classifying, or quantifying one or more DNA molecules in a DNA sample, said instructions comprising:

i.   digest said sample with said non-Type IIS restriction endonucleases into fragments, each fragment being
25 terminated on each end by a recognition site of said one or more non-Type IIS restriction endonucleases;

ii.   contact said shorter and longer strands and said digested fragments to form double stranded DNA adapters annealed to said digested fragments,

30   iii.   ligate said longer strand to said fragments such that each fragment has one said longer strand with one of said Type IIS recognition sites;

iv.   separate such of said fragments that are digested on each end, and detect a representation of the
35 length of the fragm nts and the identity of the recognition sites of said non-Typ IIS restriction ndonucleases on both termini of the fragm nts;

- 315 -

       v.   digest said fragments with one of said Type IIS restriction endonucl as s to produce second single-stranded overhangs;

       vi.   determine sequences of said second single-
5 stranded overhangs produced on said fragments by said Type IIS restriction endonucleases;

       vii.   generate one or more signals, each signal comprising said representation of the length of the fragment and the identity of effective subsequences on both termini of
10 the fragments, said effective subsequence consisting either (1) of the sequence of said non-Type IIS restriction endonuclease recognition site for the terminus not further digested by said Type IIS restriction endonuclease, or (2) of said non-Type IIS restriction endonuclease recognition site
15 combined with said sequence of said second single-stranded overhang for the terminus further digested by said Type IIS restriction endonuclease; and

       viii.   search a nucleotide sequence database to determine sequences that match or the absence of any
20 sequences that match said one or more generated signals, said database comprising a plurality of known nucleotide sequences of nucleic acids that may be present in the sample, a sequence from said database matching a generated signal when the sequence from said database has both (1) the same length
25 between occurrences of said recognition sites of said one or more non-Type IIS restriction endonucleases as is represented by the generated signal and (2) the same effective subsequences as is represented by the generated signal.

30   251.   The kit of claim 250 which comprises one or more containers having one or more double stranded adapter DNA molecules formed by annealing said longer and said shorter oligonucleotide strands.

35   252.   Th kit of claim 250 wherein said recognition sites of said Type IIS restriction endonucleases ar so positioned on said longer oligod oxynucleotide strands so

that said second single-stranded overhangs are positioned adjacent to said recognition site of said first restriction endonucleases.

5   253. A partially double stranded oligodeoxynucleotide comprising a first and a second oligodeoxynucleotide strand, said first strand oligodeoxynucleotide comprising an end complementary to a first single-stranded overhang produced by cleavage of a DNA molecule by a restriction endonuclease that

10 cuts said DNA molecule within a recognition site of said restriction endonuclease, and said partially double stranded oligodeoxynucleotide comprising a binding site for a Type IIS restriction endonuclease, said binding site so positioned with respect to

15 said end such that, when said partially double-stranded oligodeoxynucleotide is ligated to said DNA molecule cleaved by said restriction endonuclease, further cleavage by said Type IIS restriction endonuclease produces a second single-stranded overhang of said DNA molecule that is contiguous

20 with the recognition site of said restriction endonuclease.

           254. The partially double stranded oligodeoxynucleotide of claim 253 having a melting temperature no greater than 68 °C.

25

           255. The partially double stranded oligodeoxynucleotide of claim 253 wherein the melting temperature of the second oligodeoxynucleotide strand from a strand complementary to said second oligodeoxynucleotide strand is greater than the

30 melting temperature of said first oligodeoxynucleotide strand from a strand complementary to said first oligodeoxynucleotide strand.

           256. A partially double stranded oligodeoxynucleotide
35 comprising:
           (a) a first oligodeoxynucl otid strand comprising a binding means and a r lease means; and

(b)   a s cond oligodeoxynucleotide strand, said second
oligodeoxynucleotide strand comprising

(i)   a first subsequence complementary to a portion
of said first oligodeoxynucleotide strand, and

5   (ii) a second subsequence at the 5' end
complementary to a first single-stranded overhang produced by
cleavage of a DNA molecule by a restriction endonuclease that
cuts said DNA molecule within a recognition site of the
restriction endonuclease.

10

257. The partially double stranded oligodeoxynucleotide
of claim 256 wherein said release means comprises a
subsequence of said first oligodeoxynucleotide strand that is
a recognition site for a restriction endonuclease that cuts

15 rarely in the mammalian genome.

258. The partially double stranded oligodeoxynucleotide
of claim 256 wherein said release means comprises a
subsequence of one or more uracil nucleotides.

20

259. The partially double stranded oligodeoxynucleotide
of claim 256 wherein said binding means comprises a biotin
moiety attached to said first oligodeoxynucleotide strand.

25

30

35

1/25
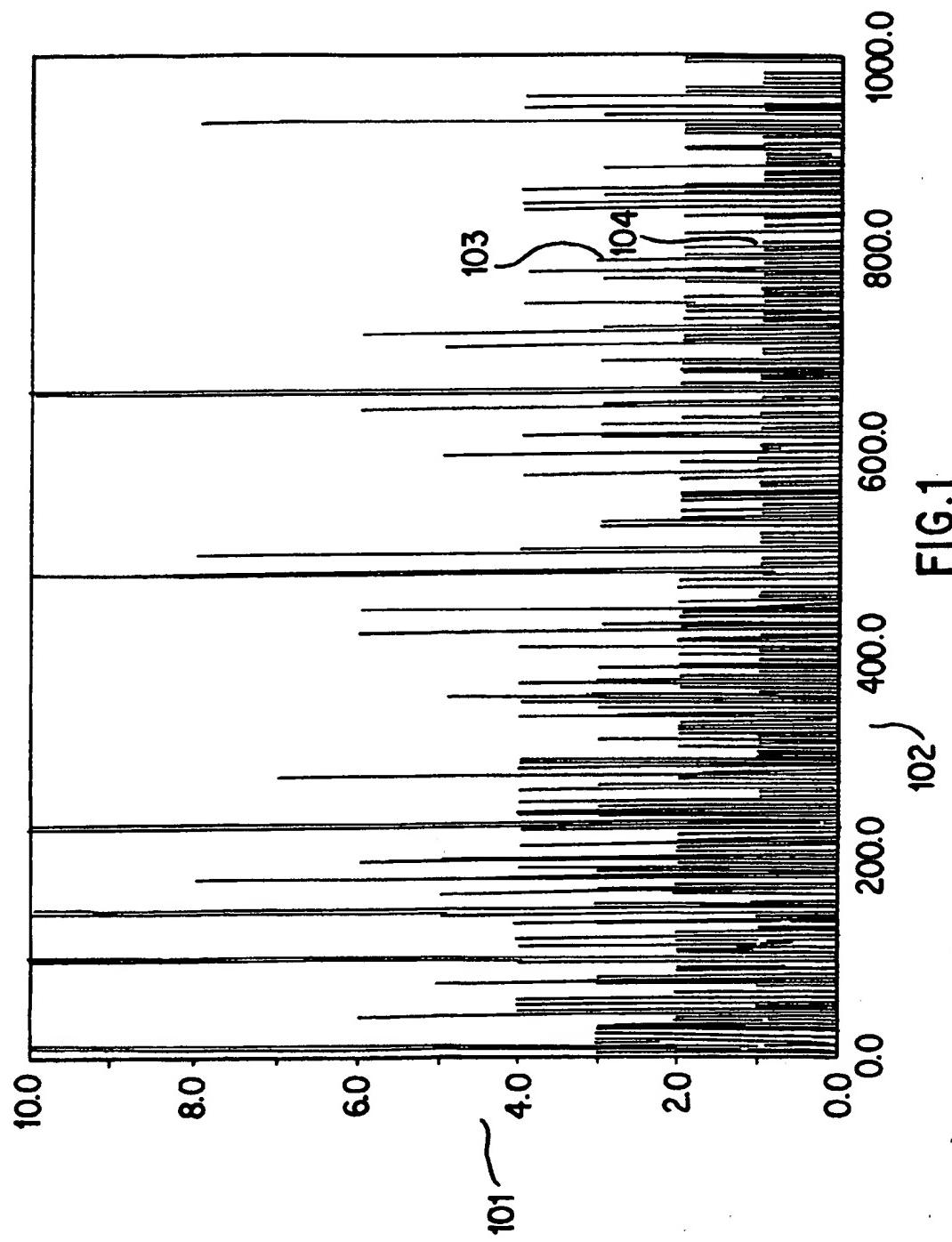
FIG.1
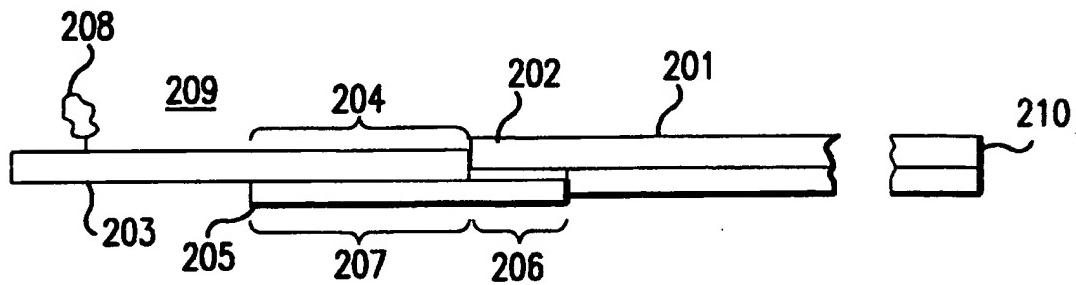
FIG.2A

5'-AGC ACT CTC CAG CCT CTC ACC GAA-3' (SEQ ID NO:1)

    250   3'- AG TGG CTT CTAG-5' (SEQ ID NO:7)

5'-ACC GAC GTC GAC TAT CCA TGA AGC-3' (SEQ ID NO:42)

    251   3'-GT ACT TCG TCGA-5' (SEQ ID NO:44)

FIG.2B
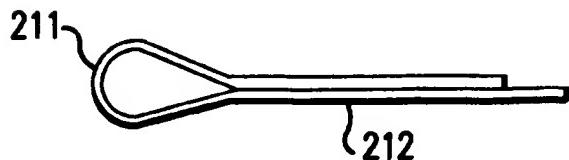
FIG.2C
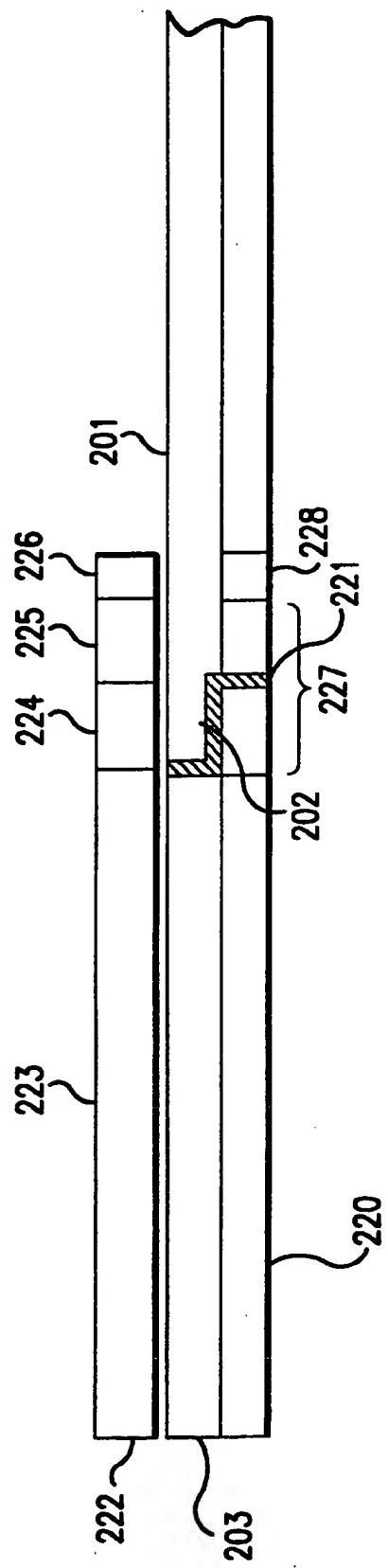
FIG.2D
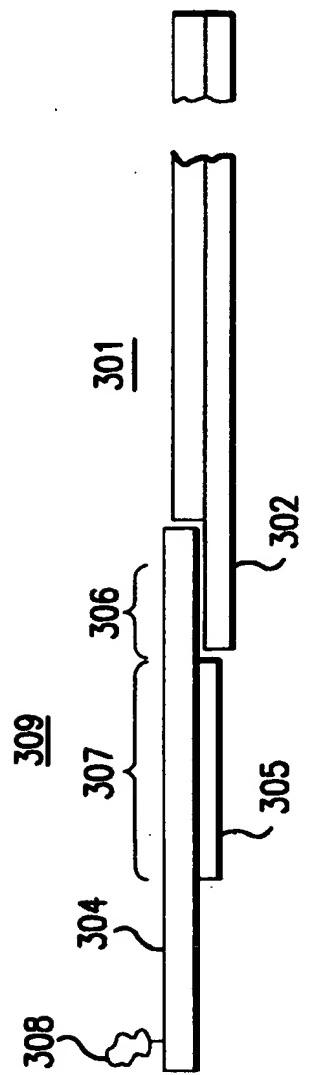
FIG.3A

5'-AGCACTCTCCAGCCTCTCACCGAGCATG (SEQ ID NO.55)
3'-AGTGGCTC
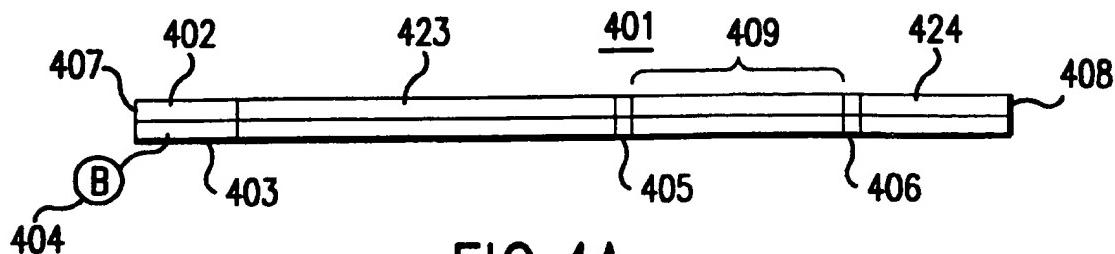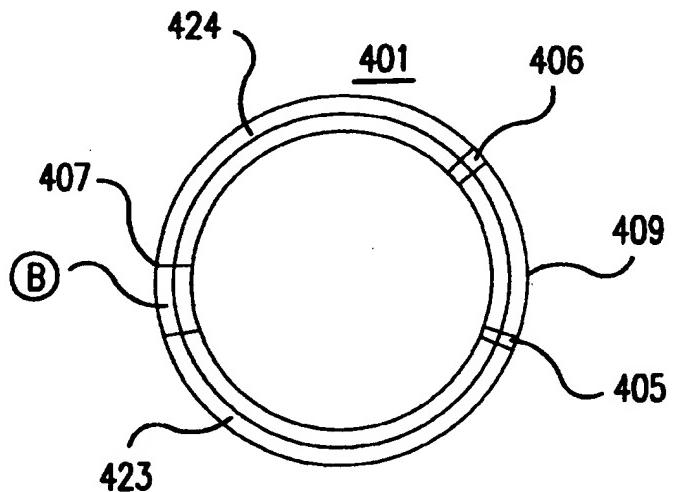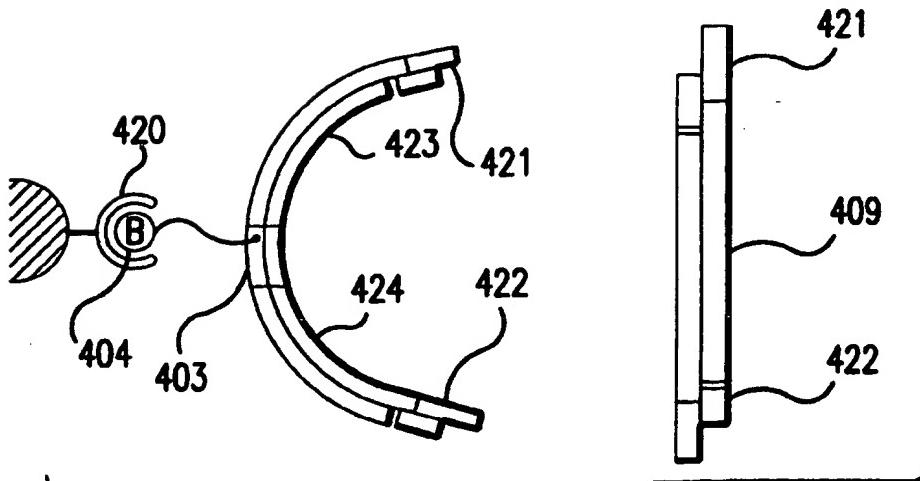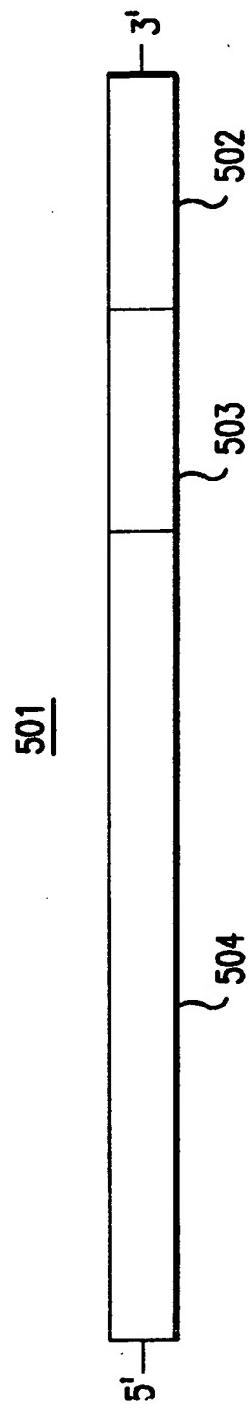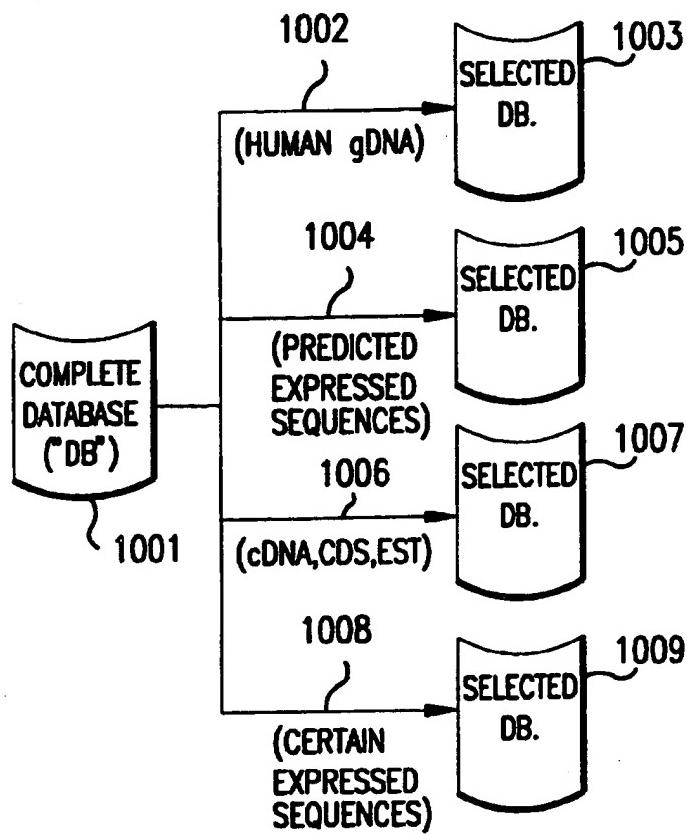
FIG.3B

FIG.4A

FIG.4B

FIG.4C

3'

502

503

501

504

FIG.5

5'

FIG.6A

1010

FIG.6B

TISSUE SAMPLE		
REACTION 1	RE1	LABEL1
	RE2	LABEL2
REACTION 2	RE1	LABEL1
	RE2	LABEL2
	RE3	LABEL3
	PROBE P	LABEL4
REACTION 15	RE36	LABEL3
	RE37	LABEL4

1101 — TISSUE SAMPLE
1102 — REACTION 1
1103 — LABEL1
1104 — LABEL2
1105 — REACTION 2
1106 — LABEL1
1107 — LABEL2
1108 — LABEL3
1112 — LABEL4
1109 — REACTION 15
1110 — LABEL3
1111 — LABEL4

1100

FIG.7

FIG.8A
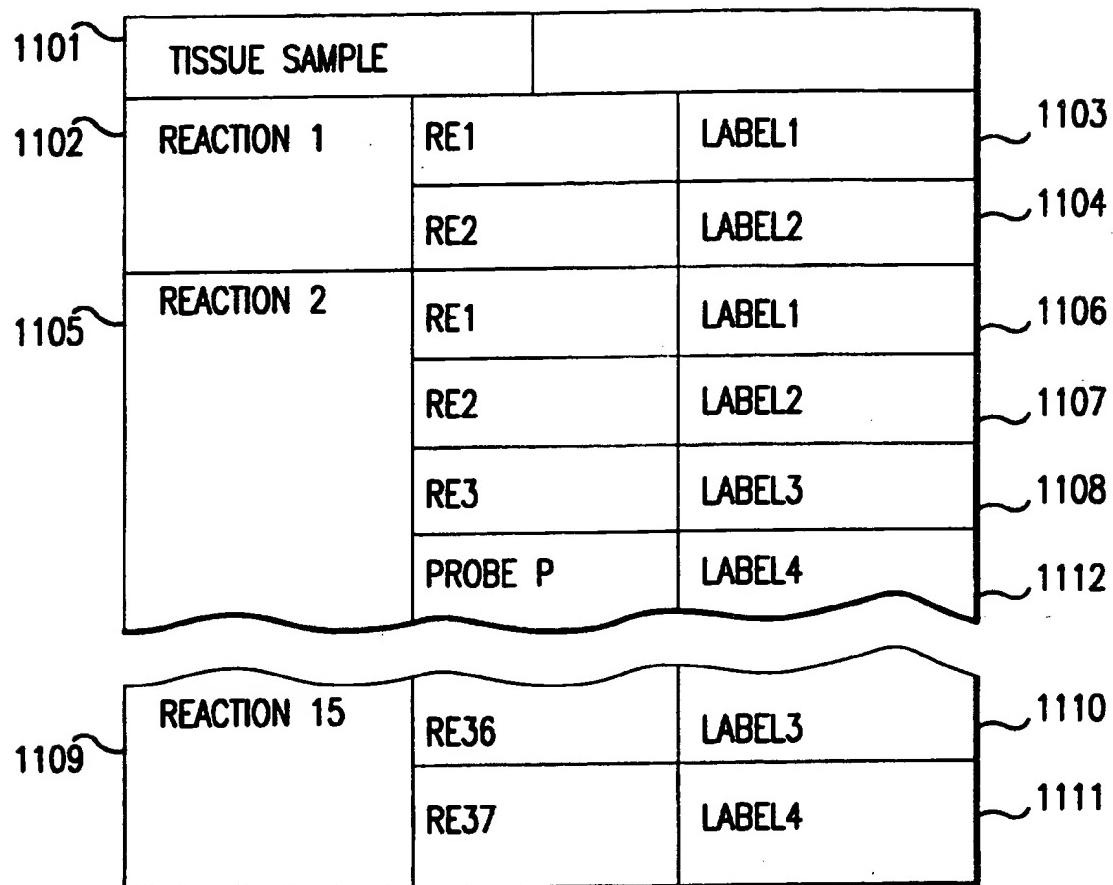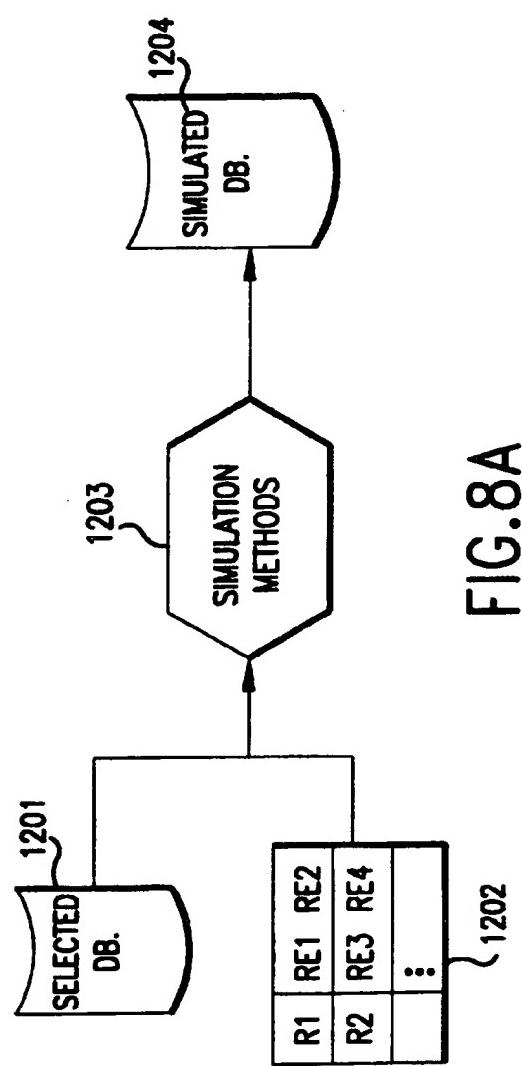
L	(RE1, RE1)R1	(RE1, RE2)R1	(RE2, RE2)R2	(RE3, RE3)R2	(RE3, RE4)R2	...
...						
52	A01			A01,S003		
...						
151	T163					
...						
175		A01,T163				
...						
222		T163,Q012				
...						
402					Q012,S003	
...						
532						
...						

1212 1213 1214 1215 1216 1217 1218 1219 1220 1211 1210

FIG.8B

START — 1301

INPUT: A SEQ.,
TARGET SUB-
SEQUENCES;
PROBE — 1302

FOR EACH SUB-
SEQ., MAKE
VECTOR OF CUTS — 1303

FOR ALL SUB-
SEQ., MERGE
AND SORT VECTORS — 1304

CREATE
FRAGMENTS
BETWEEN CUTS — 1305

PROBE
PRESENT — 1306

YES   NO

FOR EACH
FRAGMENT,
FIND ITS SEQ. — 1307

MARK IF
PROBE BINDS
TO FRAGMENT — 1308

SORT ON
FRAGMENT
LENGTH — 1309

OUTPUT:
FRAGMENT
VECTOR — 1310

END — 1311

FIG.9

1403

1401 V1 = ([10,14]$_{RE1}$,[62,66]$_{RE1}$,[610,614]$_{RE1}$)

1402 V2 = ([237,241]$_{RE2}$,[388,392]$_{RE2}$)

VECTORS OF CUTS

FIG.10A

1404 1405   1406   1407

([10,14]$_{RE1}$,[62,66]$_{RE1}$,[237,241]$_{RE2}$,

[388,392]$_{RE2}$,[610,614]$_{RE1}$)

MERGED AND SORTED VECTORS OF CUTS

FIG.10B

1408   1409

[RE1,RE1,*,52] [RE1,RE2,*,175] [RE2,RE2,*,151]

[RE1,RE2,*,222]

FRAGMENTS

FIG.10C

1410

[10,62] [62,237] [237,388] [388,610]

FRAGMENTS SEQUENCES

FIG.10D

1411
([RE1,RE1,N,52] , [RE2,RE2,y,151] ,
[RE1,RE2,N,175] , [RE1,RE2,N,222])
1412

SORTED FRAGMENT
VECTOR

FIG.10E

↕	[RE1,RE1]	[RE1,RE2]	[RE2,RE2]	
⋮				
52	+A01		1412	
⋮				
151			+A01	
⋮				
175		+A01		
⋮				
222		+A01		

DIGEST
TABLE

FIG.10F

14/25

```
                    START          1501

              ┌──────────────┐
              │ INPUT: DATA- │
              │  BASE, EXP.  │      1502
              │  DEFINITION  │
              └──────────────┘

              ┌──────────────┐
              │  INITIALIZE  │
              │    DIGEST    │      1503
              │   DATABASE   │
              └──────────────┘

┌─────────────────────────────────────────┐
│   DO  FOR  ALL  SEQUENCES  IN  DATABASE  │   1504
└─────────────────────────────────────────┘

                          ┌──────────────┐
                          │ DO NEXT MOCK │
                          │ FRAGMENTATION│   1505
                          │   REACTION   │
                          └──────────────┘

                          ┌──────────────┐
                          │    INSERT    │
                          │FRAGMENT VECTOR│  1506
                          │  IN DIGEST DB │
                          └──────────────┘

                              ╱──────────╲
                YES          ╱    MORE     ╲   1507
                            ╱   REACTIONS   ╲
                            ╲    IN  EXP.   ╱
                             ╲────────────╱
                                  NO

┌─────────────────────────────────────────┐
│                 END-DO                   │   1504
└─────────────────────────────────────────┘

              ┌──────────────┐
              │   OUTPUT:    │      1508
              │    DIGEST    │
              │   DATABASE   │
              └──────────────┘

                    END            1509
```

FIG.11

FIG.12A

FIG.12B

FIG.12C

START —— 1701

T → Ti
S → Si
E → E(S) —— 1702

1703

T≤To
OR
E≤Eo

1704 — YES

OUTPUT:
STATE,
ENERGY

NO

END —— 1705

DO N TIMES —— 1706

GENERATE
S$_{NEW}$ —— 1707

E$_{NEW}$→E(S$_{NEW}$) —— 1708

1709

E$_{NEW}$<E

YES

FIG.13A

NO

EXP[(E-E$_{NEW}$)/T]
<RANDOM NUMBER
FROM 0 TO 1 —— 1710

NO

YES

S → S$_{NEW}$
E → E$_{NEW}$ —— 1711

END-DO —— 1706

T → fT —— 1712

FIG.13B

```
                    ┌─────────┐  1801
                    │  START  │
                    └────┬────┘
                         │
                         ▼
              ┌──────────────────────┐
              │  INPUT:ACCESSION     │   1802
              │   NUMBER FOR         │
              │  SIGNAL OF INTEREST  │
              └──────────┬───────────┘
                         │
                         ▼
     ┌───────────────────────────────────────┐  1803
     │     DO FOR EACH ACCESSION NUMBER       │
     └──────┬────────────────────────┬────────┘
            │                        │
            │                        ▼
            │              ┌──────────────────┐  1804
            │              │  DO SIMULATED    │
            │              │  EXPERIMENT      │
            │              │  ON SEQUENCE     │
            │              └────────┬─────────┘
            │                       │
            │                       ▼
            │              ┌──────────────────┐  1805
            │              │     RANK         │
            │              │  SIMULATION      │
            │              │  AGAINST         │
            │              │  OBSERVATION     │
            │              └────────┬─────────┘
            │                       │
     ┌──────┴───────────────────────┴────────┐  1803
     │              END-DO                    │
     └─────────────────┬──────────────────────┘
                       │
                       ▼
              ┌──────────────────┐  1806
              │   SORT RANK      │
              │    SCORES        │
              └────────┬─────────┘
                       │
                       ▼
              ┌──────────────────┐  1807
              │   OUTPUT:        │
              │   SORTED         │
              │   SCORES         │
              └────────┬─────────┘
                       │
                       ▼
                  ┌─────────┐  1808
                  │   END   │
                  └─────────┘
```

FIG. 14

21/25

```
                    ┌─────────┐
                    │  START  │──── 1901
                    └─────────┘
                         │
                         ▼
          ┌─────────────────────┐
          │   INPUT: SET OF     │
          │  PROBE SUB-SEQ.,    │──── 1902
          │    SELECTED DB      │
          └─────────────────────┘
                         │
                         ▼
          ┌─────────────────────┐
          │     INITIALIZE      │
          │     HASH CODE       │──── 1903
          │       TABLE         │
          └─────────────────────┘
                         │
                         ▼
   ┌──────────────────────────────────────┐
   │      DO FOR SEQUENCE IN DATABASE      │──── 1904
   └──────────────────────────────────────┘
                                   │
                                   ▼
                      ┌─────────────────────┐
                      │     GENERATE        │
                      │   HASH CODE FOR     │──── 1905
                      │     SEQUENCE        │
                      └─────────────────────┘
                                   │
                                   ▼
                      ┌─────────────────────┐
                      │   ADD ACCESSION     │
                      │    NUMBER TO        │──── 1906
                      │   TABLE ENTRY       │
                      └─────────────────────┘
                                   │
                                   ▼
   ┌──────────────────────────────────────┐
   │               END-DO                  │──── 1904
   └──────────────────────────────────────┘
                         │
                         ▼
          ┌─────────────────────┐
          │     OUTPUT:         │
          │    HASH CODE        │──── 1907
          │      TABLE          │
          └─────────────────────┘
                         │
                         ▼
                    ┌─────────┐
                    │   END   │──── 1908
                    └─────────┘
```

FIG.15

65°C, 10'

37°C, 15'   37°C, 10'

−27°C, 27'

16°C, 60'

10°C, 10'   4°C, hold

FIG.16A

72°C, 10'   72°C, 10'

68°C, 1'
95°C, 30"   4°C, hold

FIG.16B

65°C, 10'

37°C, 15'   37°C, 10'

−21°C, 5'

16°C, 30'

FIG.16C

72°C, 20'   72°C, 10'

37°C, 15'   37°C, 15'

−21°C, 21'

16°C, 60'

96°C, 30'
57°C, 1'
72°C, 2"   4°C, hold

FIG.16D

FIG.17A

FIG.17B

FIG.17C

FIG.17D
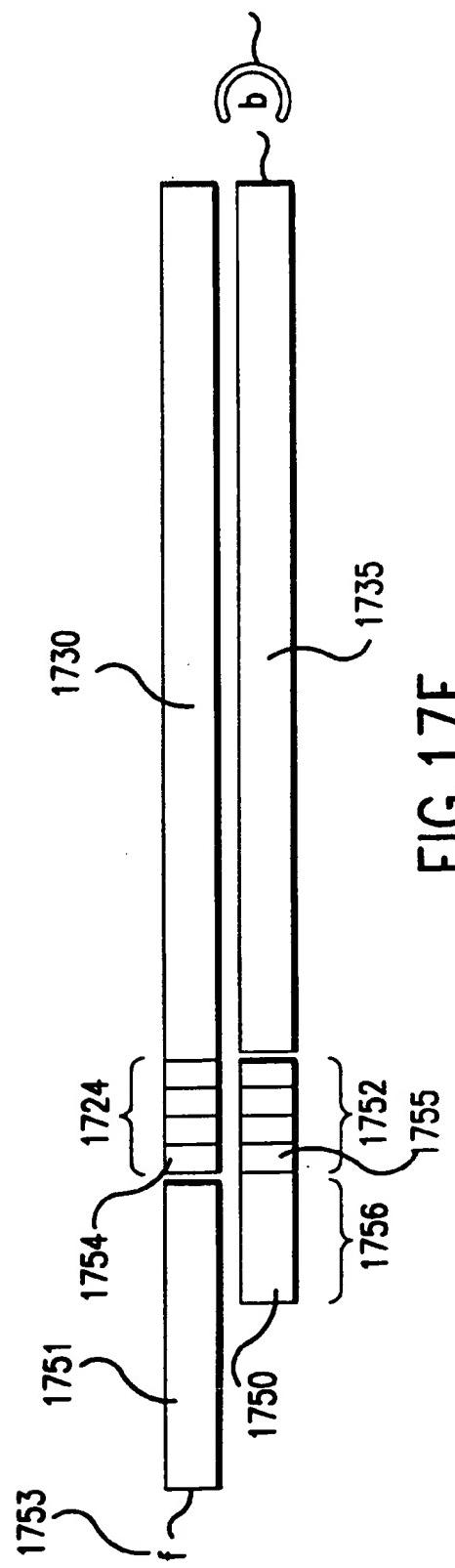
FIG.17E

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US96/17159

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) :Please See Extra Sheet.

US CL :Please See Extra Sheet.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 435/6, 5, 91.1, 91.2, 291; 422/ 82.01, 82.07; 536/ 23.1, 24.3, 24.33; 364/ 496, 499

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

Please See Extra Sheet.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	PEASE et al. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. Proc. Natl. Acad. Sci. May 1994, Volume 91, pages 5022-5026, see entire document.	1-259
Y	MEIER-EWERT et al. An automated approach to generating expressed sequence catalogues. Nature. 28 January 1993, Volume 361, pages 375-376, see entire document.	1-259
Y	MAIER et al. Hybridisation techniques on gridded high density DNA and in situ colony filters based on fluorescence detection. Nucleic Acids Research. 1994, Volume 22, Number 16. pages 3423-2424,see entire document.	1-259

| X | Further documents are listed in the continuation of Box C. | | See patent family annex. |

•	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance		
"E"	earlier document published on or after the international filing date	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed	"&"	document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
22 DECEMBER 1996	13 FEB 1997

| Name and mailing address of the ISA/US
Commissioner of Patents and Trademarks
Box PCT
Washington, D.C. 20231 | Authorized officer

DIANNE REES |
| Facsimile No. (703) 305-3230 | Telephone No. (703) 308-0196 |

Form PCT/ISA/210 (second sheet)(July 1992)*

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	ITO et al. Fluorescent differential display: arbitrarily primed RT-PCR fingerprinting on an automated DNA sequencer. FEBS Letters. 1994, Volume 351, pages 231-236, see entire document.	1-259
Y	KATO, K. Description of the entire mRNA population by a 3' end cDNA fragment generated by class IIS restriction enzymes. Nucleic Acids Research. 17 October 1995, Volume 23, Number 18, pages 3685-3690, see entire document.	1-259
Y	US 5,366,877 A (KEITH ET AL) 22 November 1994, see abstract and Figure 1.	1-259
Y	WO 95/21944 A1 (SMITHKLINE BEECHAM CORPORATION) 17 August 1995, see abstract.	137-141
Y	EIGEN et al. Sorting single molecules: Application to diagnostics and evolutionary biotechnology. Proc. Natl. Acad. Sci. USA. June 1994, Volume 91, pages 5740-5747, especially 5742-5744.	1-259
Y	JU et al. Fluorescence energy transfer dye-labeled primers for DNA sequencing and analysis. Proc. Natl. Acad. Sci. USA. May 1995, Volume 92, pages 4347-4351, see abstract.	1-259
Y	GUO et al. Direct Fluorescence analysis of genetic polymorphisms by hybridization with oligonucleotide arrays on glass supports. Nucleic Acids Research. 1994. Volume 22, Number 24, pages 5456-5465, see entire document.	1-259

INTERNATIONAL SEARCH REPORT

A. CLASSIFICATION OF SUBJECT MATTER:
IPC (6):

C12Q 1/68, 1/70; C12P 19/34; C12M 1/34; G01N 27/00, 21/64; C07H 21/02, 21/04; G06F 17/00

A. CLASSIFICATION OF SUBJECT MATTER:
US CL :

435/6, 5, 91.1, 91.2, 291; 422/ 82.01, 82.07; 536/ 23.1, 24.3, 24.33; 364/ 496, 499

B. FIELDS SEARCHED
Electronic data bases consulted (Name of data base and where practicable terms used):

APS, BIOSIS, BIOTECHABS, BIOTECHDS, BIOBUSINESS, CABA, CANCERLIT, CAPLUS, EMBASE, IFIPAT, JAPIO, MEDLINE, USPATFULL, WPIDS
search terms: gene expression , seqeunce identification, classifying, quantifying, restriction enzyme digestion, Type IIS restriction enzymes, labels, signals, fluorescent, arrays, SBH, sequencing by hybridization, mismatch detection, automation and sequencing, differential display, arrayed cDNA libraries, adoaptor or linker sequences, base calling, sequencing databases, computer programs and SBH analysis.